PSIHOLOGIE PSIHOTERAPIE

SUSANA URBINA

TESTAREA PSIHOLOGICĂ

Ghid pentru utilizarea competentă a testelor

TREI

PSIHOLOGIE · PSIHOTERAPIE

Colecție coordonată de
Simona Reghintovschi

SUSANA URBINA

Aspecte esențiale ale testării psihologice

Traducere din engleză
de Simona Reghintovschi

3
TREI

Editori:
SILVIU DRAGOMIR
VASILE DEM. ZAMFIRESCU

Director editorial:
MAGDALENA MĂRCULESCU

Director producție:
CRISTIAN CLAUDIU COBAN

Coperta:
FABER STUDIO (S. Olteanu, A. Rădulescu, D. Dumbrăvician)

Redactor:
MIRUNA ȚECHERĂ

Dtp:
OFELIA COȘMAN

Corectură:
RODICA PETCU

Descrierea CIP a Bibliotecii Naționale a României
URBINA, SUSANA
Aspecte esențiale ale testării psihologice / Susana Urbina ;
trad.: Simona Reghintovschi. - București : Editura Trei, 2009
Bibliogr.
ISBN 978-973-707-319-8

I. Reghintovschi, Simona (trad.)

159.9

Această carte a fost tradusă după:
ESSENTIAL OF PSYCHOLOGICAL TESTING, de Susana Urbina,
John Wiley & Sons, Inc., Hoboken, New Jersey, 2004

Copyright © 2004 by Susana Urbina

Copyright © Editura Trei, 2009,
pentru ediția în limba română

C.P. 27-0490, București
Tel./Fax: +4 021 300 60 90
e-mail: comenzi@edituratrei.ro
www.edituratrei.ro

ISBN 978-973-707-319-8

Cuprins

7 **Unu.** Introducere în testele psihologice și utilizarea lor

59 **Doi.** Statistica esențială pentru testare

122 **Trei.** Aspecte esențiale ale interpretării scorurilor obținute la test

184 **Patru.** Aspecte esențiale ale fidelității

237 **Cinci.** Aspecte esențiale ale validității

337 **Șase.** Considerații esențiale despre itemi

400 **Șapte.** Aspecte esențiale ale utilizării testelor

459 **Anexa A.** Testele disponibile comercial menționate în text

461 **Anexa B.** Adresele editurilor care publică teste și ale distribuitorilor

463 **Anexa C.** Tabel al ariilor situate sub curba normală

478 **Bibliografie**

502 **Mulțumiri**

Dedicaţie:
Tuturor profesorilor mei, din şcoală şi din afara şcolii, care m-au învăţat tot ceea ce ştiu, şi tuturor studenţilor mei, care m-au învăţat cum să predau.

Introducere în testele psihologice şi utilizarea lor

Primul şi cel mai general sens al termenului test înregistrat în dicţionar este „o examinare critică, observare sau evaluare". Cel mai apropiat sinonim este *probă*. La rândul său, cuvântul *critic* este definit ca „referitor la... un punct de cotitură sau un moment deosebit de important" (*Merriam-Webster's Collegiate Dictionary*, 1995). Prin urmare, nu este de mirare că, atunci când termenul *psihologic* apare după cuvântul *test*, sintagma care rezultă dobândeşte o conotaţie oarecum ameninţătoare. Testele psihologice sunt deseori utilizate pentru a evalua indivizii în anumite puncte de cotitură sau în momente semnificative din viaţa lor. Astfel, în ochii multor oameni testele par a fi probe de care depind prea mult şi despre care ei ştiu prea puţin. Într-o mare măsură, scopul acestei cărţi este de a oferi cititorilor informaţii suficiente despre testele psihologice şi despre testare pentru a înlătura conotaţiile lor ameninţătoare şi pentru a oferi mijloacele prin care utilizatorii testelor psihologice pot obţine mai multe cunoştinţe despre utilizarea lor specifică.

Mii de instrumente pot fi numite cu exactitate *teste psihologice*. Mult mai multe nu respectă această etichetă fie explicit, fie

indirect. Primul obiectiv al acestei cărți este de a explica modul în care pot fi separate primele de ultimele. Prin urmare, începem prin a defini trăsăturile pe care le au toate tipurile de teste psihologice legitime. Aceste trăsături nu doar definesc testele psihologice, dar le și diferențiază de alte tipuri de instrumente.

Testele psihologice

Un *test psihologic* este o procedură sistematică de obținere a eșantioanelor de comportament, relevante pentru funcționarea cognitivă sau afectivă și de scorare și evaluare a acestor eșantioane conform standardelor. O clasificare a fiecăruia dintre termenii principali ai acestei definiții este vitală pentru înțelegerea următoarelor discuții despre teste. Info 1.1 explică semnificația tuturor elementelor din definiția unui test psihologic și le justifică. Dacă nu sunt îndeplinite toate condițiile menționate în definiție, procedura discutată nu poate fi numită cu exactitate „test psihologic". Cu toate acestea, este important de reținut că, în esență, testele psihologice sunt pur și simplu eșantioane de comportament. Orice altceva se bazează pe inferențe.

Deseori, testele psihologice sunt descrise ca fiind *standardizate* din două motive, ambele adresându-se nevoii de obiectivitate în procesul testării. Primul are de-a face cu uniformitatea procedurii în toate aspectele importante ale administrării, scorării și interpretării testelor. Bineînțeles, momentul și locul în care este administrat un test, ca și circumstanțele în care el este administrat și examinatorul care îl administrează influențează rezultatele obținute. Cu toate acestea, scopul standardizării procedurilor de testare este de a uniformiza pe cât posibil toate variabilele aflate

Info 1.1

Elementele fundamentale ale definiției testului psihologic

Element	Explicație	Argument
Testele psihologice sunt proceduri *sistematice.*	Sunt caracterizate prin planificare, uniformitate și completitudine.	Pentru a fi utilizate, testele *trebuie* să fie obiective și corecte în mod demonstrabil.
Testele psihologice sunt *eșantioane de comportament.*	Sunt subseturi mai mici ale unui întreg mai mare.	Eșantionarea comportamentului este eficientă deoarece, de obicei, timpul disponibil este limitat.
Comportamentele eșantionate de test sunt relevante pentru *funcționarea afectivă, cognitivă* sau pentru ambele.	Eșantioanele sunt selectate pentru semnificația lor psihologică empirică sau practică.	Testele, spre deosebire de jocurile de inteligență, există pentru a fi utile; ele sunt instrumente.
Rezultatele obținute la test sunt *evaluate* și *scorate.*	Un anumit sistem numeric sau categorial este aplicat rezultatelor obținute la test, conform regulilor prestabilite.	Nu trebuie să existe nicio întrebare cu privire la ce reprezintă rezultatele obținute la test.
Pentru a evalua rezultatele obținute la test, este necesar să avem *standarde* bazate pe date empirice.	Trebuie să existe o cale de a aplica rezultatelor obținute la test o unitate de măsură obișnuită sau un criteriu.	Standardele utilizate pentru a evalua rezultatele obținute la test conduc la semnificația unică pe care o au acele rezultate.

sub controlul examinatorului, astfel încât oricui i-ar fi aplicat testul, el va fi aplicat în același mod.

Al doilea sens al standardizării privește utilizarea standardelor în evaluarea rezultatelor obținute la test. Cel mai des, aceste standarde sunt norme obținute într-un grup de indivizi — cunoscut ca grup *normativ* sau *eșantion de standardizare* — în procesul de construire a testului. Performanța colectivă a grupului sau grupurilor de standardizare, atât în termenii mediilor, cât și ai variabilității, este tabelată și devine standardul în raport cu care va fi măsurată apoi performanța celorlalți indivizi la test.

Strict vorbind, termenul *test* ar trebui să fie utilizat numai pentru acele proceduri în care răspunsurile la test sunt evaluate pe baza corectitudinii sau calității lor. Astfel de instrumente presupun întotdeauna evaluarea unui anumit aspect al funcționării cognitive al persoanei, a cunoștințelor, aptitudinilor sau abilităților sale. Pe de altă parte, instrumentele ale căror răspunsuri nu sunt nici evaluate, nici scorate drept corect-greșit sau admis-respins sunt numite *inventare, chestionare, anchete, liste, programe sau tehnici proiective* și, de obicei, sunt grupate în categoria *teste de personalitate*. Acestea sunt unelte destinate să aducă la suprafață informații despre motivațiile, preferințele, atitudinile, interesele, opțiunile, componența emoțională și reacțiile caracteristice ale unei persoane la oameni, situații și alți stimuli. De obicei, ele folosesc întrebări cu răspunsuri multiple sau de tipul adevărat-fals, cu excepția tehnicilor proiective care sunt cu final deschis. De asemenea, pot presupune alegerea forțată între afirmații reprezentând alternative contrastante sau evaluarea gradului în care cineva este sau nu este de acord cu diferite afirmații. De cele mai multe ori inventarele sau chestionarele de personalitate și alte astfel de instrumente sunt de tipul autoraportare, dar unele sunt construite și pentru a obține răspunsuri și de la alți indivizi decât persoana evaluată (ex.: de la părinte, partener sau profesor). Din rațiuni practice și urmând uzajul comun, termenul *test* va fi utilizat în această carte ca referindu-se la toate instrumentele, indiferent de tip, care se potrivesc definiției unui test psihologic. Testele care eșantionează cunoștințe, aptitudini sau funcții cognitive vor fi denumite *teste de aptitudini*, în timp ce toate celelalte vor fi denumite *teste de personalitate*.

Nu uitați!

- Cuvântul *test* are semnificație multiplă
- Termenul *test psihologic* are o semnificație foarte specifică
- În această carte, vom utiliza termenul *test* pentru a ne referi la toate instrumentele cărora li se potrivește definiția *testului psihologic*
- Ne vom referi la testele destinate eșantionării aptitudinilor, cunoștințelor sau oricărui alt tip de funcționare cognitivă cu termenul *test de aptitudini*; toate celelalte vor fi etichetate drept *teste de personalitate*

Alți termeni utilizați în legătură cu testele și denumirea testelor

Unii termeni utilizați, uneori vag, în legătură cu testele necesită explicații. Unul dintre aceștia este cuvântul *scală (scale)*, care se poate referi la:

- un întreg test alcătuit din câteva părți, de exemplu *Scala de inteligență Stanford-Binet*;
- un subtest, un set de itemi din cadrul unui test care măsoară o caracteristică distinctă și specială, de exemplu *scala*

Depresie din Inventarul Multifazic de Personalitate Minnesota (MMPI);

- o arie de subteste care au o anumită caracteristică comună, de exemplu *scalele verbale* din testele de inteligență Wechsler;
- un instrument separat alcătuit din itemi destinați să evalueze o singură caracteristică, de exemplu *Internal-External Locus of Control Scale* (Rotter, 1966);
- sistemul numeric destinat pentru a evalua sau pentru a raporta valoarea unei anumite dimensiuni măsurate, de exemplu o *scală* de la 1 la 5, cu 1 reprezentând *dezacord puternic* și 5 *acord puternic.*

Astfel, atunci când este utilizat pentru a se referi la testele psihologice, termenul *scală* devine ambiguu și lipsit de precizie. Totuși, în domeniul măsurărilor psihologice — cunoscut ca *psihometrie* — cuvântul *scală* are un sens mai precis. Se referă la un grup de itemi relevanți pentru o singură variabilă, aranjați în ordinea dificultății sau a intensității. Procesul de a ajunge la secvența de itemi este denumit *scalare.*

Baterie este un alt termen utilizat deseori în denumirile testelor. O baterie este un grup de câteva teste sau subteste administrate unei persoane la un anumit moment de timp. Atunci când câteva teste sunt asamblate de un editor pentru a fi utilizate într-un scop specific, cuvântul *baterie* apare de obicei în titlu și întregul grup de teste este văzut în totalitate, ca un singur instrument. Câteva exemple ale acestei utilizări apar în instrumentele neuropsihologice (cum ar fi Halstead-Reitan Neuropsychological Battery) în care trebuie evaluate multe funcții cognitive prin intermediul unor teste separate pentru a detecta posibile leziuni cerebrale.

Termenul *baterie* este utilizat și pentru a denumi orice grup de teste individuale selectate special de un psiholog pentru a fi utilizate cu un anumit client în efortul de a răspunde unei întrebări specifice, de obicei de natură diagnostică.

Testele psihologice ca instrumente

Fundamental pentru testele psihologice este faptul că acestea sunt instrumente. Ceea ce înseamnă că sunt întotdeauna un mijloc către un final și niciodată nu reprezintă un final în sine. Ca și alte instrumente, testele psihologice pot fi deosebit de utile — chiar de neînlocuit — atunci când sunt utilizate adecvat și cu abilitate. Cu toate acestea, testele pot fi și utilizate greșit în moduri care pot să le limiteze sau să le distorsioneze utilitatea și, uneori, chiar să ducă la consecințe dăunătoare.

O bună cale de a ilustra similaritățile dintre teste și alte instrumente mai simple este analogia dintre un test și un ciocan. Ambele sunt unelte destinate unor scopuri specifice, dar pot fi utilizate în diverse moduri. Fundamental, un ciocan este proiectat pentru bătutul cuielor în diferite suprafețe. Atunci când este utilizat adecvat, cu îndemânare, un ciocan poate ajuta la construirea unei case, asamblarea unei piese de mobilier,

Nu uitați!

Testele psihologice sunt evaluate în *două* momente distincte și prin *două* modalități distincte:

1. când sunt considerate unelte potențiale pentru viitorii utilizatori; în acest punct, principala preocupare o reprezintă calitățile lor tehnice.
2. după ce sunt utilizate pentru un anumit scop; în acest punct sunt luate în considerare în principal abilitatea utilizatorului și modalitatea în care este folosit testul.

atârnarea tablourilor într-o galerie de artă și la multe alte lucruri. Testele psihologice sunt unelte construite pentru a ajuta la realizarea inferențelor cu privire la indivizi sau grupuri. Atunci când testele sunt utilizate adecvat și cu abilitate, ele pot fi componente cheie în știința și practica psihologiei.

Așa cum ciocanele pot fi utilizate în alte scopuri bune decât cele intenționate (ex.: ca presspapier sau pentru fixarea ușii), testele psihologice pot servi și altor scopuri decât cele pentru care au fost proiectate inițial, cum ar fi sporirea auto-cunoașterii și a auto-înțelegerii. Mai mult, așa cum ciocanele pot răni oamenii și pot distruge lucrurile atunci când sunt utilizate incompetent sau malițios, testele psihologice pot fi utilizate în moduri care să producă daune. Atunci când rezultatele testului sunt interpretate greșit sau utilizate greșit, ele pot răni oamenii, etichetându-i în moduri nejustificate, refuzându-le oportunitățile în mod incorect sau, pur și simplu, descurajându-i.

Toate uneltele, fie că sunt ciocane sau teste, pot fi evaluate pornind de la cât de bine sunt proiectate și construite. Privite din acest punct de vedere, anterior utilizării, testele sunt evaluate doar într-un sens tehnic, limitat și evaluarea lor prezintă interes pentru cei mai mulți dintre utilizatorii potențiali. Odată ce ele încep să fie utilizate în practică, totuși, testele nu pot fi evaluate separat de abilitățile utilizatorilor, de modurile în care sunt utilizate și de scopurile pentru care sunt utilizate. Această evaluare în timpul utilizării presupune deseori aspecte ce țin de interese, valori sociale și chiar priorități politice. În acest context, evaluarea utilizării testelor dobândește semnificație practică pentru o mare parte a publicului.

Standardele testării

Info 1.2

Standardele testării

- această denumire va fi utilizată frecvent în această carte cu trimitere la *Standards for Educational and Psychological Testing*, publicată în 1999 de American Educational Research Association, American Psychological Association și National Council on Measurement in Education.
- *Standardele testării* reprezintă cea mai importantă sursă de criterii pentru evaluarea testelor, a practicii testării și a efectelor utilizării testelor

Datorită importanței unice a testelor pentru toți profesioniștii care le utilizează și pentru publicul general, încă de la jumătatea anilor '50, trei organizații profesionale importante și-au unit forțele pentru a promulga standarde care oferă o bază pentru evaluarea testelor, a practicii testării și a efectelor utilizării testelor. Cea mai recentă versiune este *Standards for Educational and Psychological Testing*, publicată în 1999 de American Educational Research Association (AERA) și pregătită în colaborare de AERA, American Psychological Association (APA) și National Council on Measurement in Education (NCME). Așa cum indică Info 1.2, aceste standarde sunt citate pe tot parcursul acestei cărți și de aici încolo vor fi numite *Testing Standards (Standardele testării)*.

Testele psihologice ca produse

Al doilea dintre cele mai importante fapte legate de testele psihologice este că ele sunt produse. Deși este un fapt evident, cei mai mulți oameni nu sunt conștienți de el. Testele sunt produse destinate pieței și utilizate în primul rând de psihologi profesioniști

și educatori, așa cum instrumentele stomatologice sunt prezentate pe piață și vândute stomatologilor. Publicul larg nu este conștient de natura comercială a testelor psihologice pentru că li se face reclamă în publicații și cataloage pentru publicul țintă — profesioniștii care le utilizează. Cu toate acestea, fapt este că multe, dacă nu cele mai multe, teste psihologice sunt concepute, dezvoltate, promovate pe piață și vândute în scopul de a fi aplicate în cadrul educațional, de afaceri sau în sănătatea mentală. De asemenea, ele trebuie să aducă un profit celor care le produc, ca orice alt produs comercial.

Așa cum vom vedea, încă de la început, testarea psihologică a fost alimentată în principal de nevoia de a lua decizii practice cu privire la oameni. Pentru că testele sunt instrumente profesionale care pot fi utilizate atât în beneficiul oamenilor, *cât și* ca produse comerciale, sunt justificate anumite clarificări cu privire la diferitele părți din testare și la rolul lor. Info 1.3 prezintă o listă a principalilor participanți la procesul testării și a rolurilor acestora.

Așa cum stipulează *Testing Standards* (*Standardele testării*), „interesele diferitelor părți implicate în procesul testării sunt de obicei, dar nu întotdeauna, congruente" (AERA, APA, NCME, 1999, p. 1). De exemplu, *autorii* testului sunt de obicei, dar nu întotdeauna, profesori universitari sau cercetători interesați în principal de teoretizarea sau de cercetarea psihologică, și nu de aplicațiile practice sau de profituri. *Utilizatorii* testului sunt interesați în cea mai mare măsură de adecvarea testului și de utilitatea testelor pe care le folosesc în scopuri proprii, în timp ce *editorii* sunt în mod natural înclinați să ia în considerare în primul rând profitul care poate fi obținut din vânzarea testelor. Mai mult, participanții la procesul testării pot juca unul, mai multe sau toate rolurile descrise în

Info 1.3. Utilizatorii testelor pot administra, scora și interpreta rezultatele testelor pe care le-au selectat sau pot delega una sau

Info 1.3

Participanții la procesul testării și rolurile lor

Participanți	Rolurile lor în procesul testării
Autorii testului și cei care construiesc testul	Concep, pregătesc și dezvoltă testul. De asemenea, găsesc o modalitate de a-și difuza testele, publicându-le, fie pe cale comercială, fie în publicații de specialitate cum ar fi cărțile sau revistele.
Editura care publică testul	Publică, face reclamă și vinde testele, controlând astfel distribuirea lor.
Recenzorii testului	Pregătesc critici evaluative ale testelor pe baza meritelor lor tehnice și practice.
Utilizatorii testului	Selectează sau decid să ia de pe raft un anumit test și să-l utilizeze într-un anumit scop. Ei pot participa și în alte roluri, de exemplu ca examinatori sau evaluatori.
Examinatorii sau persoanele care administrează testul	Administrează testul fie unui singur individ, fie unui grup.
Persoanele testate	Se supun testării la alegere sau din necesitate.
Persoanele care scorează testul	Calculează răspunsurile brute ale persoanei testate și le transformă în scoruri ale testului prin scorare obiectivă sau mecanică sau prin aplicarea unor judecăți de evaluare.
Persoanele care interpretează scorurile obținute la test	Interpretează rezultatele obținute la test pentru consumatori, care pot fi indivizii testați sau rudele lor, alți profesioniști sau organizații de diferite tipuri.

mai multe dintre aceste funcții altor persoane pe care le supervizează. Similar, editura care publică testul poate, și adesea se întâmplă astfel, să angajeze persoane care să construiască instrumente pentru care se consideră că există o piață. Cu toate acestea, dintre toți participanții la procesul testării, *Testing Standards* atribuie „responsabilitatea ultimă pentru utilizarea și interpretarea adecvată a testului" predominant utilizatorului testului (p. 112)

Istoria testării psihologice

Deși testele psihologice pot fi utilizate pentru a explora și investiga o largă arie de variabile psihologice, utilizarea lor fundamentală și tipică este cea de instrumente în luarea deciziilor cu privire la oameni. Nu este o coincidență faptul că testele psihologice, așa cum le cunoaștem astăzi, au luat naștere la începutul secolului XX. Anterior dezvoltării societăților urbane, industriale, democratice, exista o nevoie redusă ca numeroși oameni să ia decizii cu privire la alții, în afara celor din familiile lor sau din cercul închis al cunoștințelor. În societățile rurale, autocratice, agrare, deciziile majore de viață ce priveau indivizii erau luate, în mare măsură, de către părinți, mentori, conducători și, mai presus de toate, de genul, clasa, locul și circumstanțele în care se născuseră. Cu toate acestea, cu mult înaintea secolului XX, au existat precursori interesanți ai testării psihologice în diferite culturi și contexte.

Precursori ai testării moderne în domeniul ocupațional

O problemă perenă în orice domeniu de angajare este întrebarea cum să selectezi cel mai bine posibil oamenii pentru un loc

de muncă dat. Cei mai vechi precursori cunoscuți ai testării psihologice se găsesc tocmai în această arie, în sistemul de examinări competitive dezvoltat în Imperiul Chinez antic pentru a selecta indivizii merituoși pentru poziții de guvernare. Acest strămoș al procedurilor moderne de selecție a personalului datează din aproximativ 200 î.Hr. și a trecut printr-un număr de transformări în lunga sa istorie (Bowman, 1989). Examinările din cadrul serviciului civil chinezesc cuprindeau demonstrații ale competențelor în muzică, trasul cu arcul și călărie, printre altele, precum și examene scrise pe teme care țineau de drept, agricultură și geografie. Aparent, impulsul pentru dezvoltarea acestui sistem luminat de utilizare a resurselor umane — deschis oricărui individ care era recomandat împăratului de autoritățile locale din imperiu — era faptul că în China nu exista tipul de clasă conducătoare ereditară obișnuit în Europa până în secolul XX. Sistemul imperial chinezesc de examinare a luat sfârșit în 1905 și a fost înlocuit cu o selecție bazată pe studiile universitare. Cu toate acestea, între timp, sistemul a servit ca sursă de inspirație pentru examenele pentru serviciul civil în Britania în anii 1850 care, la rândul său, a stimulat crearea în 1860 a Civil Service Examination în Statele Unite (DuBois, 1970).

Precursori ai testării moderne în domeniul educațional

Una dintre întrebările fundamentale în orice cadru educațional se referă la modul în care studenții au achiziționat cunoștințele sau expertiza pe care profesorii lor încearcă să le transmită. Astfel, nu este surprinzător faptul că prima utilizare a testării în domeniul educațional a avut loc în timpul Evului Mediu odată cu apariția

primelor universități în Europa în secolul al XIII-lea. La acea vreme, gradele universitare au început a fi utilizate ca mijloc de a demonstra calificarea de a preda și erau planificate examinări orale formale pentru a le da candidaților la diploma universitară prilejul de a-și demonstra competența (DuBois, 1970). Încet, încet, utilizarea examinărilor s-a răspândit la al doilea nivel de educație și, pe măsură ce hârtia a devenit mai ieftină și mai disponibilă, examinările scrise au înlocuit examenele orale în cele mai multe cadre educaționale. La sfârșitul secolului al XIX-lea, atât în Europa, cât și în Statele Unite, examinările erau o metodă bine stabilită de certificare a persoanelor care trebuiau să obțină grad universitar, precum și a celor care erau capabile să exercite o anumită profesie, cum ar fi medicina sau avocatura.

Precursori ai testării moderne în psihologia clinică

O altă întrebare umană fundamentală care poate fi, și chiar a fost pusă, prin intermediul testării psihologice, este problema diferențierii „normalului" de „anormal" în arena intelectuală, emoțională și comportamentală. Cu toate acestea, în contrast cu contextele ocupaționale sau educaționale unde bazele pe care sunt luate deciziile au fost destul de clare în mod tradițional, domeniul psihopatologiei a rămas învăluit în mister și misticism pentru o perioadă mai lungă de timp.

Unii precursori ai testelor psihologice izvorăsc din domeniul psihiatriei (Bondy, 1974). Multe dintre primele teste au fost concepute în Germania în a doua jumătate a secolului al XIX-lea, deși unele dintre ele datează din prima parte a acelui secol și își au originea în Franța. Aproape invariabil, aceste instrumente erau

construite pentru scopul expres de a evalua nivelul funcționării cognitive a pacienților cu diferite tipuri de tulburări, cum ar fi retard mental sau leziuni cerebrale. Printre eșantioanele de comportament utilizate în aceste prime teste se găseau întrebări cu privire la sensul unor proverbe, diferențe și asemănări între perechi de cuvinte și sarcini de memorie precum repetarea unor serii de numere prezentate oral. Multe dintre tehnicile dezvoltate în secolul al XIX-lea erau ingenioase și au supraviețuit, fiind încorporate în testele moderne care au încă o largă utilizare (vezi McReynolds, 1986).

În ciuda isteției lor, creatorii primilor strămoși ai testelor clinice au fost dezvantajați de cel puțin doi factori. Unul era lipsa cunoștințelor — și abundența superstițiilor și concepțiilor greșite — cu privire la psihopatologie. În această privință, de exemplu, distincția dintre psihoză și retard mental nu a fost foarte clar formulată până în 1838, când psihiatrul francez Esquirol a afirmat că abilitatea de a utiliza limbajul este cel mai valid criteriu pentru stabilirea nivelului funcționării mentale a unei persoane. Al doilea factor care a împiedicat răspândirea largă și utilizarea primelor teste psihiatrice a fost lipsa lor de standardizare în termenii procedurilor sau ai unui cadru uniform de referință în raport cu care să fie interpretate rezultatele. Într-o mare măsură, tehnicile dezvoltate de neurologii și psihiatrii secolului al XIX-lea, cum ar fi Guislain, Snell, von Grashey, Rieger și alții, au fost destinate examinării unui pacient specific sau a unei anumite populații de pacienți. Aceste eșantioane de comportament erau colectate într-o manieră nesistematică și erau interpretate de clinicieni pe baza judecății lor profesionale și nu în raport cu date normative (Bondy, 1974).

Un pas semnificativ a fost realizat în psihiatrie în timpul anilor 1890 când Emil Kraepelin a prezentat o clasificare a tulburărilor

mentale în raport cu cauzele, simptomele și evoluția lor. Kraepelin dorea să aplice metoda științifică în psihiatrie și a contribuit la schițarea tabloului clinic al schizofreniei și al tulburării bipolare care — la acea vreme — erau cunoscute drept *dementia praecox*, și, respectiv, *psihoză maniaco-depresivă*. El a propus un sistem pentru a compara indivizii sănătoși și nesănătoși pe baza unor caracteristici ca distractibilitatea, senzitivitatea și capacitatea de memorare și chiar a fost pionier în utilizarea tehnicii asociației libere cu pacienți psihiatrici. Deși unii dintre studenții lui Kraepelin au construit o baterie de teste și au continuat să urmărească obiectivele stabilite de el, rezultatul muncii lor nu a fost atât de fructuos pe cât speraseră (DuBois, 1970).

Precursori ai testării moderne în psihologia științifică

Investigațiile psihofizicienilor germani Weber și Fechner de la mijlocul secolului al XIX-lea au inițiat o serie de progrese care au culminat cu crearea de către Wilhelm Wundt a primului laborator dedicat cercetărilor de natură pur psihologică la Leipzig, Germania, în 1879. Acest eveniment este considerat de către mulți drept începutul psihologiei ca disciplină distinctă formală, diferită de filosofie. Odată cu apariția noii discipline a psihologiei experimentale, a apărut și un mare interes pentru dezvoltarea aparaturii și a procedurilor standardizate pentru cartografierea ariei capacităților umane în câmpul senzației și percepției. Primii psihologi experimentaliști au fost interesați de descoperirea legilor generale care guvernează relația dintre lumea fizică și lumea psihică. Ei nu erau interesați sau erau prea puțin interesați de diferențele individuale — principalul element de interes în psihologia diferențială și testarea psihologică —, pe care, de fapt,

tindeau să le vadă ca sursă de eroare. Cu toate acestea, accentul pus pe nevoia de precizie a măsurătorilor și pe standardizarea condițiilor de laborator se va dovedi a fi o contribuție importantă la viitorul domeniu al testării psihologice.

Laboratorul german al lui Wundt a înflorit în ultimele decade ale secolului al XIX-lea și a format mulți psihologi din Statele Unite și din alte părți care s-au întors în țările lor pentru a-și construi propriile laboratoare, similare. Aproape în aceeași perioadă, un englez numit Francis Galton era interesat de măsurarea funcțiilor psihice dintr-o perspectivă total diferită. Galton era un bărbat cu o mare curiozitate intelectuală și cu multe realizări, iar poziția sa socială și financiară privilegiată i-a permis să urmărească o arie largă de interese. Era, de asemenea, văr și un mare admirator al lui Charles Darwin, a cărui teorie despre evoluția speciilor prin selecție naturală a revoluționat științele naturii la mijlocul secolului al XIX-lea. După lectura tratatului despre originea speciilor scris de vărul său, Galton a decis să-și urmărească interesul pentru ideea că dotările intelectuale tind să se moștenească în familie. În acest scop, el a stabilit în Londra un laborator antropometric în care, timp de câțiva ani, a colectat date despre câteva caracteristici fizice și fiziologice — cum ar fi lungimea brațului, înălțimea, greutatea, capacitatea vitală, forța strângerii și acuități senzoriale de diferite tipuri — de la sute de indivizi și familii. Galton era convins că aptitudinea intelectuală era o funcție a acuității simțurilor în perceperea și discriminarea stimulilor care, pe care, la rândul său, o considera ca fiind de natură ereditară. Prin acumularea și tabelarea încrucișată a datelor antropometrice, Galton spera să stabilească atât intervalul de variație al acestor caracteristici, cât și interrelațiile și concordanța între indivizi cu diferite grade de rudenie (Fancher, 1996).

Galton nu a avut succes în atingerea obiectivului său ultim, de a promova *eugenia*, un domeniu pe care el l-a înființat și care avea drept scop îmbunătățirea rasei umane prin unirea celor mai înzestrate specimene ale sale. În acest scop el dorise să inventeze o modalitate de a evalua capacitatea intelectuală a copiilor și adolescenților prin teste, astfel încât să-i identifice devreme pe cei mai dotați indivizi și să-i încurajeze să producă mai mulți urmași. Cu toate acestea, activitatea lui Galton a fost continuată și extinsă considerabil în Statele Unite de James McKeen Cattell, care a încercat, fără succes, să facă legătura între diferite modalități de măsurare a puterii discriminative, perceptive și asociative (etichetate teste „mentale") și nivelul intelectual evaluat independent, prin notele la școală.

În lumina anumitor evenimente ale secolului XX cum ar fi cele din Germania nazistă, scopul lui Galton pare o ofensă morală pentru sensibilitatea celor mai mulți contemporani. Cu toate acestea, la momentul la care el a formulat termenul *eugenie* și i-a enunțat scopurile, potențialul genocidal al acestei întreprinderi nu era general perceput și mulți indivizi iluștri ai acelor vremuri erau eugeniști entuziaști. În procesul urmăririi scopurilor sale, oricând de greșit orientat ni s-ar părea astăzi, Galton a adus contribuții importante în domeniul măsurătorilor statistice și psihologice. Realizând o cartografiere a datelor prin compararea părinților cu urmașii lor, de exemplu, el a descoperit fenomenele de regresie și de corelație, care au oferit baza multor cercetări psihologice și analize ale datelor de mai târziu. A inventat și dispozitive pentru măsurarea acuității auditive și a discriminărilor de greutate și a inițiat utilizarea chestionarelor și a asociației de cuvinte în cercetarea psihologică. Ca și cum aceste realizări nu ar fi fost suficiente, Galton a fost, de asemenea,

pionierul metodei de studiere a gemenilor care, îmbunătățită, a devenit instrumentul principal de cercetare în genetica comportamentală.

O contribuție suplimentară la proaspătul domeniu al testării psihologice la sfârșitul secolului al XIX-lea merită a fi menționată, pentru că va conduce direct la primul instrument reușit al erei moderne a testării. În timp ce studia efectele oboselii asupra aptitudinii mentale a copiilor, psihologul german Hermann Ebbinghaus — bine-cunoscut pentru importantele sale cercetări în domeniul memoriei — a pus la punct o tehnică cunoscută ca Ebbinghaus Completion Test. Această tehnică le cerea copiilor să completeze spațiile albe din pasajele unui text din care au fost omise cuvinte sau fragmente de cuvinte. Importanța acestei metode, care va fi adaptată mai târziu în diferite scopuri, este dublă. Mai întâi, pentru că a fost administrată simultan unor clase întregi de copii, a anticipat dezvoltarea testării în grup. Totuși, este mult mai important faptul că tehnica s-a dovedit a fi un instrument eficient pentru măsurarea aptitudinii intelectuale, scorurile obținute având o bună corespondență cu aptitudinea mentală a elevilor determinată de ordinea ierarhică din clasă. Drept rezultat, Alfred Binet a fost inspirat să folosească metoda completării și alte sarcini mentale complexe în dezvoltarea scalei, care va deveni primul test de inteligență reușit (DuBois, 1970).

Apariția testării psihologice moderne

La începutul anilor 1900 exista tot ce era necesar pentru apariția și dezvoltarea primelor teste psihologice cu adevărat moderne și reușite:

- Testele de laborator și instrumentele construite de primii psihologi experimentaliști din Germania
- Instrumentele de măsurare și tehnicile statistice dezvoltate de Galton și studenții săi pentru colectarea și analiza datelor privind diferențele individuale și
- acumularea unor descoperiri semnificative în științele conexe: psihologie, psihiatrie și neurologie.

Toate aceste dezvoltări au oferit fundamentul pentru apariția testării moderne. Impulsul real, cu toate acestea, a venit din nevoia practică de a lua decizii în domeniul educațional.

În 1904, psihologul francez Alfred Binet a fost repartizat într-o comisie însărcinată să construiască o metodă de evaluare a copiilor care, din cauza retardului mental sau altor întârzieri în dezvoltare, nu ar putea beneficia de orele de clasă obișnuite din sistemul public de învățământ și ar necesita educație specială. Binet era deosebit de bine pregătit pentru această sarcină, deoarece se implicase în investigarea diferențelor individuale prin intermediul diferitelor instrumente fizice și fiziologice, ca și prin teste ale proceselor mentale mai complexe, cum ar fi memoria și înțelegerea verbală. În 1905, Binet și colaboratorul său, Theodore Simon, au publicat primul instrument util pentru măsurarea aptitudinilor cognitive sau a inteligenței globale. Scala Binet-Simon din 1905, așa cum a fost cunoscut acest instrument, consta într-o serie de 30 de teste sau sarcini variate în conținut și dificultate, destinate în cea mai mare măsură evaluării judecății și capacității de raționament, fără o legătură cu învățarea școlară. Conținea întrebări referitoare la vocabular, înțelegere, diferențe între perechi de concepte și așa mai departe, precum și sarcini care includeau serii repetate de numere, urmare

a instrucțiunilor, completarea unor pasaje ale unui text fragmentat și desen.

Scala Binet-Simon a avut succes deoarece combina trăsături ale instrumentelor anterioare într-o manieră nouă și sistematică. Avea o acoperire mai mare decât instrumentele anterioare destinate evaluării unor aptitudini mai restrânse. De fapt, era o mică baterie de teste atent selectate, aranjate în ordinea dificultății și însoțite de instrucțiuni precise despre modul în care trebuie administrate și interpretate. Binet și Simon au administrat scala la 50 de copii normali, cu vârsta cuprinsă între 3 și 11 ani, și unor copii cu diferite grade de retard mental. Rezultatele acestor studii au demonstrat că ei construiseră o procedură de eșantionare a funcționării cognitive prin care nivelul general al aptitudinii intelectuale putea fi descris cantitativ, în termenii nivelului de vârstă mentală căruia îi corespundeau performanțele la scală. Nevoia pentru un astfel de instrument era atât de acută, încât scala din 1905 va fi tradusă rapid în alte limbi și adaptată pentru utilizarea în afara Franței.

Nașterea IQ

Binet însuși a revizuit, a extins și a modificat prima sa scală în 1908 și 1911. Pentru scorarea scalei a folosit un sistem în care notele pentru itemii cu răspuns corect erau acordate în termeni de ani și luni, astfel încât putea fi calculat *nivelul mental*, reprezentând calitatea performanței. În 1911, un psiholog german numit William Stern a propus ca nivelul mental atins pe scala Binet-Simon să fie reetichetat drept *scor al vârstei mentale* și să fie împărțit la vârsta cronologică a subiectului pentru a obține un coeficient mental care ar reprezenta cu mai mare exactitate aptitudinea la diferite vârste. Pentru a elimina zecimalele, coeficientul

mental a fost multiplicat cu 100 și, curând, a fost cunoscut drept *coeficient de inteligență* sau IQ. Acest scor familiar în prezent, un *raport constant IQ*, a fost popularizat prin utilizarea sa în cea mai faimoasă revizuire a scalelor Binet-Simon — Scala de inteligență Stanford-Binet — publicată în 1916 de Lewis Terman. În ciuda câtorva probleme legate de raportul IQ, utilizarea sa va dura câteva decenii, până când o modalitate mai bună de a integra vârsta și scorurile obținute la testele de inteligență (descrisă în Capitolul 3) a fost creată de David Wechsler (Kaufman, 2000; Wechsler, 1939). Ideea clasică a lui Binet — și anume că a avea o inteligență medie, peste medie și sub medie înseamnă că performanțele unei persoane la testele de inteligență se situează la nivelul tipic pentru grupul său de vârstă, deasupra sau sub acest nivel — s-a menținut și a devenit una dintre modalitățile principale prin care este evaluată inteligența.

În timp ce Binet își dezvolta scalele în Franța, în Anglia, Charles Spearman (un fost student al lui Wundt și discipol al lui Galton) a încercat să demonstreze empiric ipoteza lui Galton privind legătura dintre inteligență și acuitatea senzorială. În acest proces el a dezvoltat și a extins utilizarea metodelor de corelație ai căror pionieri au fost Galton și Karl Pearson și a oferit fundamentul conceptual pentru *analiza factorială*, o tehnică de reducere a unui număr mare de variabile la un set mai redus de factori, care vor deveni centrali pentru dezvoltarea testării psihologice și a teoriei trăsăturii.

Și Spearman a construit o teorie a inteligenței care pune accentul asupra unui factor general al inteligenței (sau *g*) prezent în toate activitățile intelectuale (Spearman, 1904a, 1904b). El a fost capabil să obțină o susținere moderată pentru ideile lui Galton, corelând evaluările realizate de profesori și notele cu rezultatele măsurătorilor acuității senzoriale, dar a realizat curând că sarcinile

cuprinse în scala Binet-Simon ofereau o modalitate mult mai utilă și mai fidelă de evaluare a inteligenței decât instrumentele folosite de el. Chiar dacă viziunile lui Spearman și Binet asupra naturii inteligenței au fost mult diferite, contribuțiile lor combinate în propulsarea dezvoltării testării psihologice în secolul XX au fost de nedepășit.

Testarea în grup

La momentul morții sale, în 1911, Binet luase deja în considerare posibilitatea de a-și adapta scala pentru alte utilizări și de a dezvolta teste de grup care să poată fi administrate de un singur examinator unor grupuri mari pentru a fi utilizate în armată și în alte situații. Totuși, concretizarea acestei idei nu va avea loc în Franța, ci în Statele Unite, unde scala Binet-Simon a fost tradusă rapid și revizuită pentru a fi utilizată în primul rând cu copii de vârstă școlară și în același scop pentru care fusese creată în Franța.

După intrarea Statelor Unite în Primul Război Modial în 1917, președintele APA, Robert Yerkes, a organizat un comitet alcătuit din psihologi pentru a ajuta la eforturile legate de război. S-a decis că cea mai practică contribuție ar fi să se construiască un test de inteligență pentru testarea în grup, care ar putea fi administrat eficient tuturor recruților din armata Statelor Unite pentru a ajuta la atribuirea sarcinilor personale. Comitetul, alcătuit din experții în testarea psihologică din acel moment, printre care Lewis Terman, a alcătuit și a probat un test care a ajuns să fie cunoscut drept *Army Alpha*. Constă din opt subteste care măsoară aptitudinile verbale, numerice, de raționament și judecată practică și informații generale. Testul, care în cele din urmă a fost administrat unui număr de peste un milion de recruți, a utilizat materiale din diferite alte instrumente, inclusiv din scalele Binet. În construirea lui, comitetul

s-a bazat puternic pe un prototip nepublicat al unui test destinat administrării în grup construit de Arthur Otis, care a imaginat itemi cu alegeri multiple ce puteau fi scorați obiectiv și rapid.

Army Alpha s-a dovedit a fi extrem de util. A fost urmat rapid de Army Beta, un test presupus a fi echivalent, care nu necesită citirea și astfel poate fi utilizat cu recruți analfabeți sau nevorbitori de limbă engleză. Din nefericire, graba cu care aceste teste au fost alcătuite și utilizate a dus la un număr de practici de testare neadecvate. În plus, au fost trase concluzii negarantate, pe baza cantităților masive de date care s-au acumulat rapid (Fancher, 1985). Unele dintre consecințele negative ale modului în care a fost implementat programul de testare în armată și ale eforturilor pentru a desfășura testarea în masă din acea epocă au deteriorat reputația testării psihologice în moduri dificil de depășit. Cu toate acestea, s-a învățat foarte mult din greșelile făcute în perioada de început a testării moderne, ceea ce mai târziu a servit la corectarea și îmbunătățirea practicilor din acest domeniu. Mai mult, odată cu testele destinate armatei, domeniul psihologiei a făcut un pas decisiv în afara laboratorului și cadrului academic și și-a demonstrat potențialul enorm de a contribui la aplicații în lumea reală.

După Primul Război Mondial, testarea psihologică și-a dobândit propriul statut în Statele Unite. Otis a publicat Group Intelligence Scale, testul care a servit drept model pentru Army Alpha în 1918, iar E. L. Thorndike, un alt pionier american important care lucra la Teachers College, Columbia, a produs în 1919 un test de inteligență pentru absolvenții de liceu, standardizat pe un eșantion selecționat (și anume, studenți în primul an). De atunci, numărul testelor publicate a crescut rapid. Rafinamente procedurale au fost, de asemenea, instituite rapid în administrarea

și scorarea testelor. De exemplu, itemii testului de diferite tipuri au început să fie prezentați într-o ordine mixtă, și nu ca subteste separate, astfel încât limita timpului global să fie utilizată pentru un test, eliminând nevoia de temporizare separată a subtestelor. Aspecte ale standardizării, cum ar fi eliminarea cuvintelor care ar putea fi citite cu diferite pronunții din testele de ortografie au trecut în prim plan, așa cum s-a întâmplat și cu *încrederea* prezentată de test — termen care, la acel moment, cuprindea ceea ce se înțelege în mod curent prin *fidelitate* și *validitate* (DuBois, 1970).

Alte progrese în domeniul testării psihologice

Succesul obținut cu testele Binet și Army a dovedit capacitatea acestora de a ajuta la luarea deciziilor cu privire la oameni. Curând, acest fapt a condus la eforturi pentru construirea unor instrumente care să ajute la luarea diferitelor tipuri de decizii. Bineînțeles, cadrele în care au apărut precursorii testelor psihologice — școli, clinici și laboratoare de psihologie — au dat naștere unor noi forme și tipuri de teste psihologice moderne.

O trecere în revistă completă a istoriei testării din prima jumătate a secolului XX depășește scopul acestei lucrări. Cu toate acestea, un rezumat al celor mai importante progrese este instructiv atât în sine, cât și pentru a ilustra diversitatea domeniului, încă din faza sa inițială.

Testarea standardizată în cadre educaționale

Pe măsură ce numărul de oameni care profită de oportunitățile educaționale la diferite niveluri a crescut, același lucru s-a întâmplat

și cu nevoia de instrumente corecte, echitabile și uniforme cu care să fie evaluați studenții în stadiile de început, de mijloc și finale ale procesului educațional. În paragrafele următoare sunt subliniate două progrese majore în testarea educațională standardizată din prima parte a secolului XX.

Testele standardizate de achiziții de cunoștințe

Introduse de E. L. Thorndike, aceste instrumente au început să evolueze din anii 1880, când Joseph Rice a demarat încercările de a studia eficiența învățării în școli. Scala de caligrafie a lui Thorndike, publicată în 1910, a deschis noi teritorii prin crearea unei serii de specimene de scris cuprinse între „foarte slab" și „excelent" cu care puteau fi comparate performanțele subiecților. Curând după aceasta, vor urma testele standardizate destinate să evalueze abilitățile aritmetice, de citire și scriere, până când măsurarea acestor subiecte sau a altora a devenit un element principal în educația elementară și secundară. Astăzi testele standardizate de achiziții de cunoștințe sunt utilizate nu numai în cadrul educațional, ci și în acordarea licențelor și certificărilor pentru profesioniștii care și-au încheiat formarea. De asemenea, ele sunt utilizate și în alte situații, printre care selecția de personal, care necesită evaluarea stăpânirii unui domeniu dat de cunoștințe.

Testele de aptitudini academice

În 1920, examinările obiective bazate pe testul Army Alpha au început să fie utilizate alături de notele din liceu în scopul luării unor decizii de admitere în colegii și universități. Această evoluție semnificativă, care a culminat cu crearea Scholastic Aptitude Test (SAT) în 1926, a anunțat apariția mult mai multor instrumente utilizate pentru selecția candidaților la școlile profesionale și

Info 1.4

Marele test

Cartea lui Nicholas Lemann (1999) *The Big Test: The Secret History of the American Meritocracy* utilizează programele de testare pentru admiterea în colegiu, în particular SAT, pentru a ilustra consecințele intenționate și neintenționate pe care astfel de programe de testare le pot avea pentru societate. Utilizarea pe scară largă a scorurilor la testele standardizate pentru luarea deciziilor sau pentru admiterea în instituții principale de învățământ superior a fost introdusă de James Bryant Conant, președintele Universității Harvard, și Henry Chauncey, primul președinte al Educational Testing Service (ETS), în anii 1940 și 1950. Obiectivul lor era să schimbe procesul prin care se obține accesul în aceste instituții — și la pozițiile de putere care, de obicei, le revin absolvenților acestor instituții — de la un proces bazat pe resurse financiare și clasă socială, la unul bazat în principal pe aptitudini demonstrate prin scorurile la teste. Lemann susține că, deși această utilizare a testării deschide porțile învățământului superior pentru copiii din clasele socio-economice medii și scăzute, generează și o nouă elită meritocratică ce se perpetuează de-a lungul generațiilor și exclude în mare măsură copiii care aparțin minorităților rasiale neprivilegiate, cărora le lipsesc oportunitățile educaționale timpurii necesare reușitei la teste.

universități. Printre cele mai cunoscute exemple ale testelor de acest tip sunt Graduate Record Exam (GRE), Medical College Admission Test (MCAT), și Law School Admission Test (LSAT), utilizate pentru programe de doctorat, facultăți de medicină și de drept. Deși fiecare dintre aceste teste conține porțiuni specifice temelor din domeniul său, în mod tipic ele au și un nucleu comun, care pune accent pe aptitudinile verbale, cantitative și de raționament necesare succesului în cele mai multe medii academice. Interesant, deși scopul lor este diferit de cel al testelor standardizate de achiziții de cunoștințe, conținutul lor este deseori similar. Info 1.4 prezintă o

relatare fascinantă a istoriei testării pentru admiterea în învățământul superior din Statele Unite.

Testarea personalului și orientarea vocațională

Utilizarea optimă a talentelor oamenilor este un obiectiv major al societății la care testarea psihologică a fost capabilă să contribuie în moduri importante aproape de la începuturile sale. Deciziile privind alegerea vocațională trebuie luate de indivizi în diferite momente ale vieții, de obicei în timpul adolescenței și tinereții dar, din ce în ce mai mult, și la vârsta mijlocie. Deciziile privind selecția și repartizarea personalului în afaceri, industrie și organizații militare trebuie luate pe o bază aflată în continuă schimbare. Unele dintre principalele instrumente care au apărut în prima perioadă și care s-au dovedit a fi deosebit de utile în luarea acestor tipuri de decizii sunt descrise în secțiunile următoare.

Teste de aptitudini și abilități specifice

Succesul testului Army Alpha a stimulat interesul pentru a selecta lucrători pentru diferite ocupații. În același timp, psihologii au elaborat și au utilizat un set fundamental de proceduri care ar justifica utilizarea testelor în selecția ocupațională. Fundamental, procedurile au presupus (a) identificarea abilităților necesare pentru un rol ocupațional dat prin intermediul unei *analize a profesiei*, (b) administrarea testelor destinate evaluării acestor abilități și (c) corelarea rezultatelor la test cu măsura performanțelor la locul de muncă. Utilizând variante ale acestei proceduri, începând din 1920, psihologii au fost capabili să construiască instrumente pentru selectarea candidaților în diferite domenii cum ar fi activități mecanice sau muzică. Testele de aptitudini

funcționărești, spațiale și motorii au urmat la scurt timp. Domeniul selecției de personal în industrie și în armată a crescut pe baza acestor instrumente, alături de utilizarea eșantioanelor de activitate, a datelor biografice și testelor de inteligență generală, individuale sau de grup. Multe dintre aceste instrumente au fost utilizate profitabil în indentificarea talentelor tinerilor care cereau orientare vocațională.

Bateriile pentru aptitudini multiple

Utilizarea testelor de aptitudini separate în consilierea vocațională va face loc într-o mare măsură în anii 1940 bateriilor pentru aptitudini multiple, construite prin tehnicile de analiză factorială inițiate de Spearman și extinse în Anglia și Statele Unite în anii 1920 și 1930. Aceste baterii sunt grupuri de teste, legate printr-un format și o bază de scorare comune, care de obicei crează profilul părților puternice și slabe ale unui individ oferind scoruri separate pentru diferiți factori cum ar fi aptitudinile verbale, numerice, spațiale, de raționament logic și mecanice, și nu un singur scor global ca IQ-ul din testele de inteligență Binet sau Army. Bateriile pentru aptitudini multiple au luat ființă în urma înțelegerii larg răspândite, datorită analizei factoriale a datelor de la testele de aptitudini, a faptului că inteligența nu este un concept unitar și că aptitudinile umane cuprind un spectru larg de componente sau factori separați și relativ independenți.

Instrumente pentru măsurarea intereselor

Așa cum testele de abilități speciale și de aptitudini au apărut în industrie și, mai târziu, au fost utilizate în consilierea vocațională, instrumentele pentru măsurarea intereselor au fost destinate inițial orientării vocaționale și, mai târziu, au fost utilizate

în selecția de personal. Truman L. Kelley a creat în 1914 Interest Test, un test simplu, probabil primul inventar de iterese cu itemi care privesc preferințele de lectură și activități din timpul liber, precum și itemi care presupun cunoașterea cuvintelor și informații generale. Cu toate acestea, principala descoperire în această arie specifică a avut loc în 1924, când M. J. Ream a construit o grilă de scorare empirică, ce diferenția răspunsurile vânzătorilor cu succes și fără succes la Carnegie Interest Inventory, construit de Yroakum și studenții săi de la Carnegie Institute of Technology în 1921 (DuBois, 1970). Acest eveniment a marcat începuturile unei tehnici cunoscute drept *criteriu empiric de scorare* care, după modificări cum ar fi procedurile de validare încrucișată și extinderi la alte profesii, va fi utilizată în Strong Vocational Interest Blank (SVIB), publicat inițial în 1927, și în alte tipuri de inventare. Versiunea prezentă a SVIB — numită Strong Interest Inventory® (SII) — reprezintă unul dintre cele mai larg utilizate inventare de interese căruia i s-au alăturat multe alte instrumente de acest tip.

Testarea clinică

La începutul secolului XX, domeniul psihiatriei s-a orientat către modalități mai sistematice de clasificare și studiere a psihopatologiei. Aceste evoluții au oferit impulsul pentru dezvoltarea instrumentelor care ar putea ajuta la diagnosticarea problemelor psihiatrice. Principalele exemple ale acestui tip de instrumente sunt discutate în continuare.

Inventarele de personalitate

Primul instrument de acest tip a fost Woodworth Personal Data Sheet (P-D Sheet), un chestionar construit în timpul Primului Război Mondial pentru a tria recruții care ar putea suferi de boli mentale. Era alcătuit din 116 afirmații privind sentimentele, atitudinile și comportamentele care indicau în mod evident psihopatologia și la care se răspundea simplu cu „da" sau „nu". Deși P-D Sheet prezenta anumite promisiuni, Primul Război Mondial s-a sfârșit înainte ca instrumentul să intre în utilizarea operațională. După război a urmat o perioadă de experimentare cu alte tipuri de itemi, mai puțin evidenți, și cu scale desemnate să evalueze nevroticismul, trăsăturile de personalitate — cum ar fi introversia și extraversia — și valorile. Au apărut inovații în prezentarea itemilor care au avut drept scop reducerea influenței dezirabilității sociale, cum ar fi tehnica alegerii forțate introdusă în Allport-Vernon Study of Values în 1931. Cu toate acestea, inventarul de personalitate care a avut cel mai mare succes în acea perioadă și care mai supraviețuiește și astăzi a fost Inventarul Multifazic de Personalitate Minnesota (MMPI; Hathaway & McKinley, 1940). MMPI a combinat itemi din P-D Sheet și alte inventare, dar a utilizat tehnica de scorare bazată pe criteriu empiric introdusă odată cu SVIB. Această tehnică a dus la un instrument mai puțin transparent, la care respondenții nu puteau să disimuleze atât de ușor, pentru că mulți dintre itemi nu făceau o referire evidentă la tendințele psihopatologice.

Încă din 1940, inventarele de personalitate au înflorit. Au fost introduse multe modificări în construirea lor, inclusiv utilizarea unor perspective teoretice — cum ar fi sistemul de trebuințe al lui Murray (1938) — și a metodelor de consistență internă în selectarea itemilor. În plus, analiza factorială, care a avut o importanță atât de mare în studierea și diferențierea aptitudinilor, a început să fie utilizată în construirea inventarelor de personalitate. În anii 1930, J. P. Guilford a fost pionierul utilizării

analizei factoriale la itemii grupați în scale omogene, în timp ce, în 1940, R. B. Cattell a aplicat această tehnică, încercând să identifice trăsăturile centrale de personalitate și, prin urmare, merită să fie investigate și evaluate. În prezent, analiza factorială joacă un rol important în cele mai multe aspecte ale teoriei testării și teoriei construirii testului.

Tehnicile proiective

Deși inventarele de personalitate au avut un anumit succes, profesioniștii din sănătatea mentală care lucrau cu populații de pacienți de psihiatrie au simțit nevoia unui ajutor suplimentar în diagnosticarea și tratarea bolilor mentale. În anii 1920, a apărut un nou gen de instrumente pentru evaluarea personalității și psihopatologiei. Aceste instrumente, cunoscute ca *tehnici proiective*, își aveau rădăcinile în metoda asociației libere al cărei pionier a fost Galton și care a fost utilizată clinic de Kraepelin, Jung și Freud. În 1921 un psihiatru elvețian, Hermann Rorschach, a publicat un test alcătuit din 10 planșe cu pete de cerneală, care erau prezentate examinatului pe rând, pentru a fi interpretate. Cheia succesului acestui prim test proiectiv formal a fost că a oferit metode standardizate de obținere și interpretare a răspunsurilor subiecților la planșe, răspunsuri care — în mare — reflectau modurile unice de percepție ale subiectului și de relaționare a sa cu lumea. Tehnica lui Rorschach a fost preluată de psihologii americani și s-a propagat în diferite universități și clinici din Statele Unite după moartea sa timpurie din 1922. Tehnica Rorschach, alături de alte instrumente care se bazează pe imagini, cuvinte și desen, cum ar fi Testul de Apercepție Tematică (TAT), testele de completare a frazelor și desenarea unei figuri umane, a oferit un întreg nou repertoriu de instrumente — mai subtile și mai incisive decât

chestionarele — cu care clinicienii puteau investiga aspecte ale personalității pe care cei testați ar fi incapabili sau nu ar fi dispuși să le dezvăluie. Deși există mari controverse cu privire la validitatea lor, în primul rând pentru că se bazează deseori pe interpretări calitative la fel de mult sau chiar mai mult decât pe scorurile numerice, tehnicile proiective constituie încă o parte semnificativă a instrumentarului multor clinicieni. (Viglione & Rivera, 2003).

Testele neuropsihologice

Rolul disfuncțiilor cerebrale în tulburările emoționale, cognitive și comportamentale a fost recunoscut din ce în ce mai mult de-a lungul secolului trecut. Cu toate acestea, impulsul major pentru studiul clinic și științific al relațiilor creier-comportament, care constituie subiectul *neuropsihologiei*, a provenit din investigațiile lui Kurt Goldstein asupra dificultăților pe care le-a observat la soldații care au suferit leziuni cerebrale în timpul Primului Război Mondial. Deseori, acești soldați prezentau un tipar de deficiențe care includea probleme în gândirea abstractă, memorie și planificarea și executarea unor sarcini relativ simple, toate fiind cunoscute sub rubrica de *organicitate* care a fost utilizată ca sinonim pentru leziunile cerebrale. După câteva decenii, au luat naștere numeroase instrumente menite să detecteze organicitatea și să o distingă de alte tulburări psihiatrice. Multe dintre acestea erau variante ale testelor de performanță — ca opuse testelor verbale — care fuseseră construite pentru a evalua aptitudinea intelectuală generală a indivizilor care nu puteau fi examinați în limba engleză sau care prezentau deficiențe de vorbire sau auz. Aceste teste au utilizat materiale ca planșe, puzzle-uri și cuburi și sarcini creion-hârtie, cum ar fi labirinturi și desene. În ultimele câteva decenii

s-au aflat multe despre creier și funcționarea sa și a fost necesar să se revizuiască multe dintre primele concepții cu privire la evaluarea neuropsihiatrică pe baza noilor informații. Leziunea cerebrală nu mai era privită drept o condiție de tipul totul sau nimic a organicității, cu un set comun de simptome, ci ca o arie largă de posibile tulburări rezultate din interacțiunea factorilor genetici și comportamentali specifici fiecărui caz individual. Cu toate acestea, numărul și tipurile de instrumente disponibile din domeniul evaluării neuropsihologice au continuat să crească și au contribuit la înțelegerea clinică și științifică a numeroaselor și variatelor relații dintre funcționarea cerebrală și cogniție, emoții și comportamente. (Lezak, 1995).

Utilizarea curentă a testelor psihologice

Testarea contemporană este, în general, mai sofisticată din punct de vedere metodologic și mai bine informată decât până acum. Utilizarea actuală a testelor are loc în diferite contexte și poate fi clasificată în trei categorii: (a) luarea deciziei, (b) cercetarea psihologică și (c) înțelegerea de sine și dezvoltarea personală. Așa cum se poate observa din lista prezentată în Info 1.5, cele trei tipuri de utilizări diferă mult în

Info 1.5

Utilizarea curentă a testelor psihologice

- Prima și cea mai importantă utilizare a testelor este în procesul pragmatic de *a lua decizii cu privire la oameni*, atât indivizi, cât și grupuri.
- A doua utilizare a testelor în termenii frecvenței și longevității este *cercetarea științifică a fenomenelor psihice* și a diferențelor individuale.
- Cea mai recentă și cea mai puțin dezvoltată utilizare a testelor este în procesul terapeutic *de promovare a înțelegerii de sine și a ajustării psihologice.*

privința impactului pe care îl au și a multor alte privințe, dar prima utilizare este cea mai vizibilă pentru publicul larg.

Luarea deciziilor

Utilizarea principală a testelor psihologice este cea de instrumente pentru luarea deciziei. Această aplicație specifică a testelor psihologice presupune invariabil judecăți de valoare din partea uneia sau mai multor persoane care iau decizii și care trebuie să determine bazele pe care să selecteze, să repartizeze, să clasifice, să diagnosticheze sau să aibă de-a face în alt fel cu indivizi, grupuri, organizații sau programe. Bineînțeles, această utilizare a testării este deseori plină de controverse, deoarece duce deseori la consecințe nefavorabile pentru una sau mai multe părți. În multe situații în care testele sunt utilizate pentru a lua decizii, iar oamenii nu sunt de acord cu deciziile luate, utilizarea testelor este atacată indiferent dacă a fost sau nu adecvată.

Atunci când testele sunt utilizate pentru a lua decizii semnificative cu privire la indivizi sau programe, testarea ar trebui să fie o parte a unei strategii de luare a deciziei completă și bine planificată, care ia în considerare contextul particular în care sunt luate deciziile, limitele testelor și alte surse de date alăturate testelor. Din nefericire, foarte des — din motive de operativitate, lipsă de grijă sau de informație — testele sunt făcute să poarte responsabilitatea proceselor defectuoase de luare a deciziilor, care pun prea multă greutate asupra rezultatelor la teste și neglijează alte informații relevante. Numeroase decizii luate de instituții guvernamentale, de învățământ sau de corporații pe o bază rutinieră, presupunând de obicei evaluarea simultană a mai multor oameni în același timp, au fost și sunt încă luate în această

manieră. Deşi ele au consecinţe importante pentru indivizii implicaţi — cum ar fi angajarea, admiterea la colegii sau şcoli de formare profesională, absolvire sau certificare pentru practicarea unei profesii — deciziile se bazează aproape exclusiv pe scorurile la teste. Profesioniştii din testare şi unele instanţe guvernamentale încearcă să schimbe această practică, o moştenire a modului în care a apărut testarea. Unul dintre cei mai importanţi paşi în această direcţie este publicarea unui ghid al utilizării testelor ca parte a luării deciziilor deosebit de importante cu privire la studenţi destinat educatorilor şi celor care stabilesc politica organizaţională (U.S. Department of Education, Office for Civil Rights, 2000).

Cercetarea psihologică

Testele sunt deseori utilizate în cercetările din domeniul psihologiei diferenţiale, educaţionale, sociale, vocaţionale, din psihologia dezvoltării sau psihopatologie, printre altele. Ele oferă o metodă recunoscută de studiere a naturii, dezvoltării şi interrelaţionării trăsăturilor cognitive, afective şi comportamentale. De fapt, deşi câteva teste care îşi au originea în desfăşurarea investigaţiilor psihologice au ajuns să fie disponibile pe piaţă, mult mai multe instrumente sunt arhivate în disertaţii, reviste şi diferite compendii ale instrumentelor experimentale prezentate la finalul acestui capitol în Surse de informaţii despre teste. Deoarece rareori există consecinţe practice imediate ale utilizării testelor în cercetare, utilizarea lor în acest context este mai puţin discutabilă decât atunci când testele sunt utilizate în luarea deciziilor cu privire la indivizi, grupuri, organizaţii sau programe.

Cei mai mulţi psihologi şi consilieri cu orientare umanistă au perceput, deseori justificat, că domeniul testării accentuează excesiv etichetarea şi clasificarea indivizilor în termenii unor criterii numerice rigide. Începând din anii 1970, câţiva dintre ei, în mod remarcabil Constance Fischer (1985/1994), au început să utilizeze testele şi alte instrumente de evaluare într-o manieră individualizată, concordantă cu principiile psiihologiei umaniste şi existenţial-fenomenologice. Această practică ce priveşte testarea ca un mod de a oferi clienţilor informaţii în scopul promovării înţelegerii de sine şi a dezvoltării personale pozitive a dus la *modelul terapeutic de evaluare* susţinut de Finn şi Tonsager (1997). În mod evident, aplicarea acestui model are cea mai mare relevanţă în cadre de consiliere şi psihoterapie în care clientul este singurul şi principalul utilizator al rezultatelor testului.

Evaluarea psihologică versus testarea psihologică

Din motive legate în cea mai mare parte de comercializarea testelor, unii autori şi unele edituri au început să utilizeze cuvântul *evaluare* în titlurile testelor. Astfel, în mintea publicului general termenii *evaluare* (assessment) şi *testare* (testing) sunt consideraţi deseori ca sinonimi. Aceasta este o evoluţie nefericită. Numeroşi profesionişti din domeniu consideră că distincţia dintre aceşti termeni merită păstrată şi că publicul general, potenţialii clienţi ai evaluării sau consumatori ai testelor, trebuie să fie conştient de această distincţie.

Utilizarea testelor pentru luarea deciziilor cu privire la o persoană, un grup sau un program ar trebui să aibă loc întotdeauna

Nu uitați!

- *Testele* și *evaluările* NU sunt sinonime
- Testele se numără printre instrumentele utilizate în procesul de evaluare

în contextul unei *evaluări psihologice*. Acest proces poate avea loc în domeniul asistenței sănătății, consilierii sau în cel judiciar, precum și în domeniul educațional sau ocupațional. Evaluarea psihologică este un *proces* flexibil, nestandardizat, care are scopul de a ajunge la o concluzie justificată cu privire la unul sau mai multe aspecte sau întrebări psihologice prin colectarea, evaluarea și analiza datelor conform scopului urmărit (Maloney & Ward, 1976).

Pași ai procesului de evaluare

Primul și cel mai important pas în evaluarea psihologică este de a identifica scopurile sale pe cât de clar și de realist posibil. Fără obiective clar definite în privința cărora evaluatorul și persoana care solicită testarea sunt de acord, nu este probabil ca procesul să fie satisfăcător. În cele mai multe cazuri procesul de evaluare se finalizează cu un raport scris sau verbal, prin care se comunică persoanelor care au solicitat testarea concluziile la care s-a ajuns într-o manieră comprehensibilă și utilă. Între aceste două puncte, specialistul care a condus evaluarea, de obicei un psiholog sau un consilier, va avea nevoie să facă apel la expertiza sa la fiecare pas. Acești pași presupun selectarea adecvată a instrumentelor care urmează să fie utilizate în colectarea datelor, administrarea lor atentă, scorarea, interpretarea și — cea mai importantă dintre toate — utilizarea judicioasă a datelor colectate pentru a face inferențe cu privire la întrebarea în cauză. Acest ultim pas trece dincolo de expertiza în psihometrie și necesită

cunoștințe în domeniul legat de întrebare, cum ar fi, printre altele, îngrijirea sănătății, repartizarea educațională, psihopatologie, comportament organizațional sau criminologie. Exemple de aspecte ce pot fi investigate prin evaluare psihologică includ:

- *Întrebări legate de diagnostic*, cum ar fi diferența dintre depresie și demență;
- *Realizarea predicțiilor*, cum ar fi estimarea probabilității comportamentelor suicidare sau omucidare;
- *Judecăți evaluative*, cum sunt cele care apar în deciziile privind custodia unui copil sau în evaluarea eficienței unor programe sau intervenții.

Niciunul dintre aceste aspecte complexe nu poate fi rezolvat doar prin intermediul scorurilor la test, pentru că același scor la test poate avea semnificații diferite în funcție de persoana examinată și de contextul în care a fost obținut. Mai mult, niciun scor sau set de scoruri nu poate surprinde toate aspectele care trebuie luate în considerare în soluționarea unor astfel de probleme.

Testele psihologice pot fi componentele cheie ale evaluării psihologice, dar testarea și evaluarea diferă fundamental în moduri importante. Info 1.6 prezintă câteva dimensiuni care diferențiază testarea și evaluarea psihologică. Deși există puține întrebări cu privire la superioritatea generală a evaluării asupra testării cu privire la comprehensivitate și utilitate, cu cât este mai mare complexitatea procesului de evaluare, cu atât rezultatele sale sunt mai dificil de apreciat decât cele ale testării. Cu toate acestea, în ultimii ani, dovezile eficacității evaluării, cel puțin în domeniul acordării îngrijirii sănătății, au început să fie reunite (Eisman et al., 2000; Kubiszyn et al., 2000; Meyer et al., 2001).

Info 1.6

Diferențe tipice între testarea și evaluarea psihologică

Baza	Testarea psihologică	Evaluarea psihologică
Gradul de complexitate	Mai simplă; presupune o procedură uniformă, frecvent unidimensională	Mai complexă, fiecare evaluare presupune diferite proceduri (intverviu, observație, testare, etc.) și dimensiuni
Durata	Mai scurtă, cu o durată cuprinsă între câteva minute și câteva ore	Mai lungă, durând de la câteva ore la câteva zile sau mai mult
Sursele de date	O persoană, cel căruia i se aplică testul	Deseori sunt folosite surse colaterale, pe lângă subiectul evaluat, cum ar fi rude sau profesori
Focalizarea	Modul în care o persoană sau un grup de persoane este în comparație cu alții (nomotetic)	Unicitatea unui individ dat, a unui grup sau a unei situații (ideografică)
Calificări necesare utilizării	Cunoașterea testelor și a procedurilor de testare	Cunoștințe despre testare și alte metode de evaluare, dar și despre aria evaluată (ex.: tulburări mentale, cerințele postului)
Baza procedurală	Este necesară obiectivitatea; cuantificarea este deosebit de importantă	Este necesară subiectivitatea sub forma judecății clinice; cuantificarea este rareori posibilă
Costuri	Necostisitoare, mai ales atunci când testarea se realizează în grup	Foarte costisitoare; necesită utilizarea intensivă a unor profesioniști înalt calificați

Baza	Testarea psihologică	Evaluarea psihologică
Scop	Obținerea datelor ce vor fi utilizate în luarea deciziilor	Decizie care privește întrebarea sau problema solicitantului
Grad de structură	Înalt structurată	Conține aspecte structurate și nestructurate
Evaluarea rezultatelor	Investigații relativ simple ale validității și fidelității pe baza rezultatelor obținute în grup	Foarte dificilă datorită variabilității metodelor, evaluatorilor, naturii întrebărilor prezentate etc.

Calificarea utilizatorului de teste

Pe măsură ce numărul testelor a continuat să crească și utilizarea lor s-a extins nu doar în Statele Unite, ci în întreaga lume, problema utilizării greșite a testelor a devenit o preocupare din ce în ce mai mare pentru public, guvern și diferite profesii. Psihologia, profesia din care iau naștere testele și cea cu care acestea sunt cel mai limpede asociate, a preluat conducerea în încercarea de a împiedica utilizarea lor greșită. *Testing Standards*, promulgată de APA și alte organizații profesionale (AERA, APA, & NCME, 1999), reprezintă un mijloc major al acestui scop. De asemenea, APA s-a adresat și aspectelor legate de testare și evaluare în principiile sale etice și în codul de conduită (APA, 2002), ca și alte asociații profesionale (ex.: American Counseling Association, 1995; National Association of School Psychologists, 2000).

Deși calitățile tehnice ale unor teste sunt departe de a fi ideale și pot contribui la probleme ce apar în utilizarea lor, se consideră, în general, că motivul principal al utilizării greșite a testelor se

află în cunoştinţele sau competenţele insuficiente ale numeroşilor utilizatori ai testelor. Testele pot părea relativ simple şi uşor de înţeles pentru potenţialii utilizatori care nu sunt conştienţi de precauţiile care trebuie luate în aplicarea lor. Din acest motiv, în ultimele decenii, asociaţiile profesionale din Statele Unite şi din alte părţi au construit instrumente care subliniază mai clar şi mai specific decât înainte abilităţile şi cunoştinţele de bază necesare pentru utilizarea competentă (American Association for Counseling and Development, 1988; Eyde, Moreland, Robertson, Primoff, & Most, 1988; International Test Commission, 2000; Joint Committee on Testing Practices, 1988).

Una dintre cele mai clare expuneri ale acestor cerinţe este un raport pregătit pe parcursul a cinci ani de Task Force on Test User Qualifications (APA, 2000). Acest raport subliniază (a) cunoştinţele şi abilităţile centrale esenţiale ale celor care utilizează testul pentru a lua decizii sau a formula politici care afectează viaţa subiecţilor testaţi şi (b) expertiza pe care trebuie să o deţină utilizatorii testului în contextul specific al recrutării de personal, educaţiei, consilierii carierei, îngrijirii sănătăţii şi activităţii judiciare. Cunoştinţele şi abilităţile centrale sau generale în psihometrie, statistică, selectarea testelor, administrare, scorare, întocmirea raportului şi protejarea rezultatelor sunt considerate relevante pentru toţi utilizatorii testului. Cunoştinţe suplimentare şi experienţa supervizată necesare utilizării testelor în diferite contexte şi cu diferite grupuri de subiecţi sunt, de asemenea, subliniate în raport, ca şi diferite utilizări ale testelor în clasificarea, descrierea, predicţia, planificarea intervenţiilor şi urmărirea lor în fiecare dintre aceste contexte.

Un alt aspect al testării care a contribuit la utilizarea greşită a testelor de-a lungul anilor este relativa uşurinţă cu care instrumentele de testare pot fi obţinute de persoane care nu sunt calificate pentru a le utiliza. Într-o anumită măsură, disponibilitatea testelor depinde de libertatea cu care circulă informaţia în societăţi democratice ca Statele Unite, în special în era internetului. Un alt motiv al acestei probleme — la care am făcut aluzie mai devreme în acest capitol — este faptul că multe teste sunt produse comerciale. Ca rezultat, unele edituri sunt dispuse să vândă teste unor persoane sau instituţii fără a utiliza măsuri adecvate de protecţie pentru a stabili dacă acele persoane deţin acreditările potrivite. La un anumit moment, în timpul anilor 1950 şi 1960, *Testing Standards* a inclus un sistem tristratificat de clasificare a testelor în termenii calificărilor necesare utilizării lor (APA, 1966, pp. 10–11). Acest sistem, care etichetează testele ca aparţinând nivelului A, B sau C în funcţie de formarea necesară pentru a le utiliza, a fost uşor evitat de indivizii din şcoli, agenţii guvernamentale şi afaceri. Deşi multe edituri folosesc încă acest sistem, *Testing Standards* nu o mai face. Info 1.7 subliniază elementele incluse în mod tipic în sistemul de clasificare tristratificat al calificărilor utilizatorului de teste.

În 1992, câteva edituri care publică teste şi furnizori de servicii în domeniul testării au înfiinţat Association of Test Publishers (ATP). Această organizaţie nonprofit încearcă să menţină un nivel înalt al profesionalismului şi eticii în testarea psihologică. O cale prin care monitorizează distribuirea testelor este cererea unei documentaţii care să ateste un nivel minim de formare al celor care cumpără produsele.

Formularele de calificare pentru cumpărarea testului sunt incluse acum în cataloagele tuturor editurilor renumite. Indiferent cât de sincere pot fi editurile în eforturile lor de a asigura securitatea materialelor testului şi a preveni utilizarea lor greşită, eficienţa acestor eforturi este limitată de necesitate. Nu numai că

nu este fezabil să verifici calificările pe care cumpărătorii susţin că le au în formularele trimise, dar, mai mult, niciun set formal de calificări — obţinute printr-o formă de învăţământ sau prin licenţiere — nu poate asigura faptul că individul este competent pentru a utiliza adecvat un anumit test într-o situaţie dată (vezi Capitolul 7).

Surse de informaţii despre teste

În domeniul testării psihologice, ca şi în orice alt domeniu al activităţii umane, internetul a creat un stoc de informaţii inepuizabil. Astfel, alături de referinţele tipărite care existau în mod tradiţional în domeniu, există acum un mare număr de resurse elecronice şi media online uşor accesibile.

Info 1.7

Nivelele de calificare ale utilizatorilor de teste psihologice

Toate editurile renumite care publică teste le cer cumpărătorilor să completeze un formular în care să specifice acreditările care îi califică pentru a utiliza materialele testelor pe care doresc să le cumpere şi care certifică faptul că materialele vor fi utilizate în acord cu toate reglementările etice şi legale valabile. Deşi numărul de nivele şi de acreditări specifice necesare pentru fiecare nivel diferă de la o editură la alta, de obicei criteriile de calificare sunt organizate pe cel puţin trei nivele, bazate aproximativ pe o clasificare a testelor şi cerinţelor de formare subliniate iniţial de American Psychological Association (APA; 1953, 1954)

	Cel mai jos strat (Nivelul A)	Strat intermediar (Nivelul B)	Cel mai înalt strat (Nivelul C)
Tipul de instrumente la	O gamă limitată de instrumente,	Instrumente care cer o anumită	Instrumente care cer o familiari-

	Cel mai jos strat (Nivelul A)	Strat intermediar (Nivelul B)	Cel mai inalt strat (Nivelul C)
care se referă acest nivel	cum ar fi teste de achiziţii educaţionale care pot fi administrate, scorate şi interpretate fără o formare specializată, urmând instrucţiunile din manual.	formare specializată în construirea şi utilizarea testelor şi în domeniul în care aceste instrumente vor fi aplicate, cum ar fi testele de aptitudini şi inventarele de personalitate aplicabile populaţiei normale.	zare extinsă cu principiile testării şi evaluării, ca şi cu domeniile psihologice la care se referă instrumentele, cum ar fi testele individuale de inteligenţă şi tehnicile proiective.
Tipuri de acreditări sau cerinţe necesare cumpărării materialelor la acest nivel	Unele edituri nu cer nicio acreditare pentru cumpărarea testelor de la acest nivel. Altele pot cere studii universitare într-un domeniu adecvat, comenzi de materiale realizate prin intermediul unei agenţii sau instituţii sau ambele.	De obicei cumpărătorii de teste trebuie să aibă un nivel de master în psihologie (sau într-un domeniu conex) sau o activitate continuă în testare şi evaluare proporţională cu cerinţele pentru utilizarea instrumentului respectiv.	Cumpărătorii de teste trebuie să aibă tipul de formare avansată şi experienţă supervizată dobândită pe parcursul obţinerii gradului de doctor, o licenţă profesională într-un domeniu relevant pentru utilizarea instrumentului sau ambele.

Resurse pe internet

Pentru persoanele care caută informații despre testele psihologice, un bun punct de plecare este secțiunea de Testare și Evaluare a site-ului web APA (http://www.apa.org). În cadrul acestei secțiuni există un excelent articol despre „Întrebări frecvente/Aflarea informațiilor despre testele psihologice" (APA, 2003) care oferă un ghid al modului în care pot fi localizate teste publicate și nepublicate și documente importante relevante pentru testarea psihologică. *Testele publicate* sunt comercializate de edituri care publică teste, deși uneori, ca și cărțile, ajung în situația de a nu mai fi disponibile în stoc. *Testele nepublicate* trebuie să fie obținute direct de la cercetătorul care le-a creat, cu excepția situației în care au apărut în reviste de specialitate sau în publicații speciale (prezentate pe scurt).

Două alte mari portaluri de pe internet pentru cei care caută informații despre un anumit test sunt (a) Buros Institute of Mental Measurements (BI) pagina web Test Reviews Online (http://www.unl.edu/buros), care oferă informații gratuite despre aproape 4 000 de teste disponibile pe piață și peste 2 000 de recenzii de teste care pot fi cumpărate și afișate online; și (b) Educational Testing Service (ETS) Test Collection database (http://www.ets.org/testcoll/index.html), care este una dintre cele mai mari de acest tip din lume. În plus, sistemul site-ului web Educational

Nu uitați!

Una dintre distincțiile fundamentale dintre teste se referă la publicarea lor.

- *Testele publicate* sunt comercializate de edituri care publică teste.
- *Testele nepublicate* trebuie obținute de la cercetătorii care le-au construit, din indexurile speciale ale instrumentelor nepublicate sau din revistele de specialitate.

Resources Information Center (ERIC) (http://eric.ed.gov) — fondat de U.S. Department of Education — conține numeroase materiale legate de testarea psihologică.

O altă cale de a obține informații online despre testele publicate și nepublicate este utilizarea indexărilor revistelor de psihologie, educație sau afaceri. Baza de date PsycINFO de la APA, disponibilă în multe biblioteci sau pe bază de abonament, oferă un portal în care, folosind denumirea testului, se pot găsi referințe bibliografice, rezumate și chiar texte complete ale articolelor despre acesta. În PsycINFO și alte baze de date se poate căuta după subiect, cuvinte cheie și autori, ceea ce le face deosebit de utile atunci când sunt disponibile doar informații parțiale.

Nu uitați!

Anexa A enumeră toate testele și instrumentele de evaluare psihologică publicate, disponibile comercial și menționate în această carte, împreună cu codurile de identificare ale editurilor care le-au publicat.

Anexa B oferă adrese de internet actuale ale editurilor enumerate în Anexa A. Informații mai detaliate despre editurile care publică teste, inclusiv adrese și numere de telefon, sunt disponibile în ultima ediție *Tests in Print* (Murphy, Plake, Impara & Spies, 2002).

Din momentul în care un test este localizat prin intermediul oricăreia dintre aceste resurse, de obicei se poate afla dacă este publicat și cum poate fi obținut. Dacă testul este publicat, el poate fi comandat de la compania care îl publică de către cei care au calificările de a-l utiliza. Informația necesară comenzii poate fi găsită în catalogul editurii, multe cataloage fiind acum disponibile atât online, cât și în formă tipărită. Site-ul ATP (http://www.testpublishers.org) are legături către multe edituri care publică teste și oferă servicii de evaluare psihologică. Adresele de internet ale tuturor organizațiilor menționate în această

secțiune și alte surse importante de informații despre teste pot fi găsite în Info 1.8.

Resurse tipărite

Teste publicate

Referitor la testele publicate, disponibile comercial, cele mai importante surse de informații provin de la Buros Institute of

Info 1.8

Surse de informații pe internet despre testele psihologice

Organizație (acronim)	Website
American Educational Research Association (AERA)	http://www.aera.net
American Psychological Association (APA)	http://www.apa.org
Association of Test Publishers (ATP)	http://www.testpublishers.org
Buros Institute of Mental Measurements (BI)	http://www.unl.edu/buros
Educational Resources Information Center (ERIC)	http://eric.ed.gov
Educational Testing Service (ETS)	http://www.ets.org/testcoll/index.html
International Test Commission (ITC)	http://www.intestcom.org
National Council on Measurement in Education (NCME)	http://www.ncme.org

Mental Measurements (BI) din Lincoln, Nebraska. În particular, BI (http://www.unl.edu/ buros) produce două serii de volume care pot ghida utilizatorii cu privire la aproape orice test publicat disponibil în Statele Unite. Unul dintre acestea este seria *Tests in Print* (*TIP*), iar celălalt este seria *Mental Measurements Yearbook* (*MMY*). *Tests in Print* este o bibliografie completă a tuturor testelor disponibile comercial în momentul în care un anumit volum al seriei este publicat. Fiecare intrare conține titlul testului, acronimul, autorul, editura, data publicării și alte informații de bază despre test, precum și referințe ale revizuirilor testului din toate volumele *MMY* disponibile în acel moment. Suplimentar, seria *TIP* conține un index clasificat extrem de util al testelor care sunt *tipărite*, ca și indexuri ale scorurilor testelor, edituri, acronime și nume ale autorilor și recenzenților. La rândul său, seria *MMY* apare din 1938, când regretatul Oscar Buros a publicat primul anuar pentru a ajuta utilizatorii de teste, furnizând recenzii evaluative ale testelor scrise de profesioniști independenți și calificați. Deși *MMY* este publicat încă sub formă de carte, intrările și recenziile sunt, de asemenea, disponibile online și pe suport elecronic. Buros Institute publică și multe alte materiale legate de teste.

PRO-ED (http://www.proedinc.com) publică *Tests*, o serie de volume enciclopedice care prezintă scurte descrieri ale instrumentelor din psihologie, educație și afaceri. Seria *Test Critiques*, care datează din 1984, acompaniază *Tests*. Fiecare volum din această serie conține recenzii ale testelor și indexuri cumulative ale tuturor volumelor sale anterioare.

Teste nepublicate

Obiectivul oamenilor de știință comportamentaliști care utilizează teste psihologice este de a investiga constructele psihologice

și diferențele individuale și de grup. Multe teste existente sunt utilizate exclusiv în cercetarea științifică și nu sunt disponibile comercial. Aceste teste sunt denumite instrumente *nepublicate*, deoarece ele nu pot fi cumpărate; de obicei, condițiile pentru utilizarea lor sunt stabilite de autorii fiecărui instrument și deseori este nevoie de o scrisoare prin care se cere permisiunea de a le utiliza. Informații despre teste nepublicate — și deseori chiar instrumentele — pot fi găsite în revistele de psihologie (ex.: prin PsycINFO online) și în diferite indexuri (ex.: Goldman, Mitchell, & Egelson, 1997; Robinson, Shaver, & Wrightsman, 1991). Articolul menționat anterior „Întrebări frecvente/Găsirea informațiilor despre testele psihologice" (APA, 2003) enumeră câteva resurse electronice și tipărite de informații cu privire la teste nepublicate.

TESTEAZĂ-TE

1. **Care dintre următoarele afirmații *nu* este un element esențial al testării psihologice?**
   (a) Proceduri sistematice
   (b) Utilizarea unor standarde obținute empiric
   (c) Reguli de scorare prestabilite
   (d) Eșantionarea comportamentului afectiv

2. **Cea mai importantă sursă a criteriilor de evaluare a testelor, aplicării testelor și efectelor testelor poate fi găsită în**
   (a) *Ethical Principles of Psychologists and Code of Conduct.*
   (b) *Standards for Educational and Psychological Testing.*
   (c) *Diagnostic and Statistical Manual of Mental Disorders.*
   (d) *Report of the Task Force on Test User Qualifications.*

3. **Primii precursori ai testării moderne pentru selecția de personal datează din**
   (a) China, Î.Hr.
   (b) Grecia antică.
   (c) Imperiul incaș.
   (d) Europa medievală.

4. **Evaluarea testelor psihologice este *cel mai puțin* problematică**
   (a) anterior punerii lor în utilizare.
   (b) odată ce au intrat în uz.

5. **Comparativ cu alte domenii enumerate, dezvoltarea criteriilor sau a bazelor pentru luarea deciziei a fost substanțial mai încetinită în contextul**
   (a) evaluării educaționale.
   (b) evaluării ocupaționale.
   (c) evaluării clinice.

6. **Meritele construirii primului test psihologic reuşit în era modernă îi sunt atribuite în general lui**
   (a) Francis Galton.
   (b) Alfred Binet.
   (c) James McKeen Cattell.
   (d) Wilhelm Wundt.

7. **Raportul constant IQ sau coeficientul de inteligenţă a fost obţinut prin**
   (a) adunarea vârstei mentale (VM) şi a vârstei cronologice (VC) a celui testat.
   (b) scăderea VC din VM şi înmulţirea rezultatului cu 100.
   (c) împărţirea VC la VM şi înmulţirea rezultatului cu 100.
   (d) împărţirea VM la VC şi înmulţirea rezultatului cu 100.

8. **Scopul principal pentru care testele psihologice sunt folosite în mod curent este**
   (a) cercetarea psihologică.
   (b) cercetarea educaţională.
   (c) luarea deciziilor.
   (d) înţelegerea de sine şi dezvoltarea personală.

9. **Comparativ cu testarea psihologică, evaluarea psihologică este în general**
   (a) mai simplă.
   (b) mai structurată.
   (c) mai costisitoare.
   (d) mai obiectivă.

10. **Care dintre următoarele va fi cea mai bună sursă de informaţii despre un test care *nu* este disponibil comercial?**
    (a) *Mental Measurements Yearbooks*
    (b) *Test Critiques*
    (c) *Tests in Print*
    (d) *PsycINFO*

Răspunsuri corecte: 1. d; 2. b; 3. a; 4. a; 5. c; 6. b; 7. d; 8. c; 9. c; 10. d.

Statistica esenţială pentru testare

În general, progresul ştiinţei se îmbină cu inventarea instrumentelor de măsură şi cu progresele metodelor şi tehnicilor de măsurare. De exemplu, ştiinţa astronomiei a luat cu adevărat avânt în secolele al XVII-lea şi al XVIII-lea în urma inventării telescopului potrivit pentru observarea cosmosului şi a inventării de către Descartes a geometriei analitice care, printre altele, a condus la o calculare mai precisă a distanţelor dintre corpurile cereşti. În mod similar, paşii decisivi făcuţi în prezent în domeniul neuroştiinţelor datorează mult dezvoltării unor tehnici ca tomografia prin emisie de pozitroni (PET) şi imagistica funcţională prin rezonanţă magnetică (fMRI) care permite oamenilor de ştiinţă să vizualizeze şi să măsoare micile schimbări şi evenimente biochimice din creier.

Aşa cum am văzut în Capitolul 1, şi domeniul modern al testării psihologice a început cu inventarea unor instrumente reuşite. Scalele de inteligenţă Binet-Simon destinate măsurării proceselor cognitive importante — cum ar fi înţelegerea, raţionamentul şi memoria — prin eşantioane de comportament calibrate conform vârstei. Inventarea de către Arthur Otis a itemilor obiectivi cu

alegere multiplă a condus la primele testări de grup ale inteligenței generale. Tehnicile statistice dezvoltate aproximativ în același timp cu primele teste au permis analiza datelor colectate prin intermediul acelor teste.

Măsurarea

Conceptul de măsurare se află în centrul testării psihologice ca întreprindere științifică de studiere a comportamentului uman. *Măsurarea* presupune utilizarea anumitor instrumente sau a unor reguli de atribuire de numere unor obiecte sau evenimente (Stevens, 1946). Dacă aplicăm acest proces în mod sistematic, atunci, într-o mai mare măsură și cu mai mare ușurință, un fenomen măsurat este supus confirmării și analizei și, astfel, devine mai obiectiv. Cu alte cuvinte, prin analiza sistematică, clasificarea și cuantificarea fenomenelor observabile, plasăm aceste fenomene în arena științifică.

Pentru definiția testelor psihologice este central faptul că ele constau în eșantioane de comportament alese cu atenție cărora li se aplică un sistem numeric sau categorial conform unor standarde prestabilite. Testarea psihologică este în mare măsură coextensivă cu aria *psihometriei* sau a măsurării psihologice și reprezintă unul dintre instrumentele primare ale teoriei și practicii în psihologie.

Utilizarea numerelor în testare ne cere să ne adâncim în statistică. Pentru mulți studenți la psihologie utilizarea statisticii și a datelor cantitative în general constituie o problemă ce poate părea insurmontabilă: a avea de-a face cu numere tinde să trezească o oarecare anxietate. Această anxietate este legată de distresul pe care cursurile de matematică și statistică îl induc deseori datorită unor motive ce pot fi legate la fel de mult atât de factori

emoționali și atitudinali ai subiecților, cât și de modul în care aceste materii sunt predate în mod tradițional. Acest capitol prezintă conceptele statistice necesare pentru înțelegerea principiilor fundamentale ale testării psihologice. Cei care stăpânesc noțiunile fundamentale de statistică pot trece peste cea mai mare parte a capitolului. Cât despre ceilalți, orice cititor motivat al acestei cărți poate ajunge să înțeleagă conceptele descrise aici. Totuși, este important să realizați că aceste concepte urmează o desfășurare logică; pentru a trece la o temă nouă este esențial să le stăpâniți pe cele precedente. Un ajutor suplimentar în înțelegerea metodelor statistice fundamentale poate fi găsit cu ușurință în multe manuale excelente, cum sunt cele enumerate în Info 2.1.

Variabile și constante

Una dintre distincțiile de bază făcute în orice știință este cea dintre variabile și constante. Așa cum presupun chiar termenii, o *variabilă* este ceva ce variază, în timp ce o *constantă* este ceva ce nu variază. Lumea noastră are multe variabile și puține constante. Un exemplu de constantă este π (pi), raportul dintre circumferința cercului și diametrul său, un număr care este de obicei rotunjit la 3,1416. Pe de altă parte, variabilele se află pretutindeni și pot fi clasificate în numeroase modalități. De exemplu, unele variabile sunt vizibile (sex, culoarea ochilor) și altele invizibile (personalitate, inteligență); unele sunt definite astfel încât să se refere la seturi foarte reduse, iar altele la seturi foarte mari (numărul de copii dintr-o familie sau venitul mediu al indivizilor dintr-o țară); unele sunt continue, iar altele discrete.

Ultima distincție este importantă pentru scopurile noastre și merită anumite explicații. Din punct de vedere tehnic, variabilele

Info 2.1

Recomandări cu privire la statistică

Premise fundamentale

1. Pentru a înțelege testele psihologice, trebuie să lucrezi cu numere și statistică.
2. Înțelegerea statisticii este posibilă pentru oricine citește această carte
3. Cea mai bună cale de a înțelege mai bine conceptele statistice este aplicarea lor.

Surse de ajutor recomandate

Cărți

- Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.
- Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Fort Worth, TX: Harcourt Brace.
- Urdan, T. C. (2001). *Statistics in plain English.* Mahwah, NJ: Erlbaum.
- Vogt, W. P. (1998). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (2nd ed.).Thousand Oaks, CA: Sage.

Materiale video

- Blatt, J. (Producer/Writer/Director). (1989). *Against all odds: Inside statistics* [VHS]. (Disponibilă la The Annenberg/CPB Project, 901 E St., NW, Washington, DC 20004–2006)

discrete sunt cele cu un interval finit de valori — sau cu un interval de valori potențial infinit, dar numărabil. De exemplu, variabilele *dihotomice* sunt variabile discrete care pot lua doar două valori, cum ar fi sexul sau fețele unei monede. Variabilele *politomice* sunt variabile discrete care pot lua mai mult de două valori, cum ar fi starea civilă, rasa și așa mai departe. Alte variabile discrete pot lua mai multe valori, dar pot fi încă numărate ca unități separate; astfel de exemple sunt mărimea familiei, numărul vehiculelor din

trafic și scorurile de baseball. Deși este posibil ca în practică să faci erori de numărare, în principiu variabilele discrete pot fi înregistrate (calculate) precis și fără eroare.

Variabile *continue* cum ar fi timpul, distanța și temperatura, pe de altă parte, au intervale infinite și nu pot fi numărate. Ele sunt măsurate cu scale care, teoretic, pot fi subdivizate la infinit și nu au pauze între punctele lor, cum ar fi scalele din ceasurile analogice, metrul de lemn și termometrele din sticlă. Din moment ce instrumentele noastre de măsură (chiar și ceasurile atomice!) nu pot fi niciodată calibrate cu suficientă precizie pentru a măsura cu exactitate variabile continue, măsurătorile pe care le facem acestor variabile sunt aproximări mai mult sau mai puțin exacte.

Înainte de a începe să avem de-a face cu numere, este necesar un avertisment. În testarea psihologică suntem aproape întotdeauna interesați de variabile continue (ex.: gradul de integritate, extraversie sau anxietate), însă măsurăm cu instrumente cum ar fi testele sau inventarele care nu sunt la fel de precise ca cele din științele fizice și biologice. Chiar și în aceste științe măsurarea discretă a variabilelor continue pune anumite limite preciziei

Nu uitați!

- Deși numerele pot părea precise, toate măsurătorile sunt predispuse erorilor.
- Când măsurăm variabile discrete, erorile apar numai din numărarea imprecisă. O bună practică cere prevenirea, detectarea și corectarea numărării imprecise.
- Atunci când măsurăm variabile continue, pe de altă parte, eroarea de măsură este inevitabilă, fiind o consecință a limitelor instrumentelor de măsură.
- Ca instrumente de măsură, testele psihologice sunt supuse multor limite. Astfel, intervalul de eroare trebuie să fie întotdeauna estimat și comunicat împreună cu rezultatele testării.

măsurătorilor. Prin urmare, este evident că în ştiinţele comportamentului trebuie să fim conştienţi în special de potenţialele surse de eroare şi să căutăm estimări pertinente ale erorilor ori de câte ori ne sunt prezentate rezultatele unui proces de măsurare. De exemplu, dacă sondajele realizate pe eşantioane de potenţiali votanţi sunt utilizate pentru a estima rezultatele alegerilor, limitele estimate ale erorii trebuie să fie prezentate alături de rezultatele sondajului.

Rezumând, atunci când privim rezultatele oricărui proces de măsurare, trebuie să păstrăm clar în minte faptul că sunt *inexacte*. În special în testarea psihologică, ori de câte ori sunt prezentate scorurile obţinute la un test, trebuie clarificat faptul că acestea sunt estimări; mai mult, limitele în care scorurile pot varia şi nivelele de încredere pentru acele limite trebuie precizate, alături de informaţiile furnizate de interpretarea testului (vezi Capitolul 4).

Semnificaţia numerelor

Deoarece numerele pot fi folosite în diferite moduri, S. S. Stevens (1946) a construit un sistem de clasificare a diferitelor nivele de măsurare pe baza relaţiilor dintre numere şi obiectele sau evenimentele cărora li se aplică numerele. Aceste nivele de măsurare sau scale — prezentate în Tabelul 2.1 — specifică unele dintre diferenţele majore ale modalităţilor în care numerele pot fi utilizate şi tipurile de operaţii statistice care sunt admise logic în funcţie de modul în care sunt utilizate numerele.

Tabelul 2.1 Nivele de măsură

Tipul de scală	Caracteristici definitorii	Proprietăţile numerelor	Exemple
Nominală	Numerele sunt folosite în locul cuvintelor	Identitate sau egalitate	Numerele de securitate socială (SS#), numerele de pe tricourile jucătorilor de fotbal; coduri numerice pentru variabile necantitative, cum ar fi sexul sau diagnosticul psihiatric
Ordinală	Numerele sunt utilizate pentru a ordona serii ierarhice	Identitate + ordinea rangului	Ordonarea după rang a atleţilor sau a echipelor; percentile
Interval	Intervale egale între unităţi, dar fără zero adevărat	Identitate+ ordinea rangului+ egalitatea unităţilor	Temperatura pe scalele Fahrenheit şi Celsius; timpul din calendar
Raport	Zero înseamnă „fără" ceea ce este măsurat; toate operaţiile aritmetice sunt posibile *şi* au sens.	Identitate+ ordinea rangului+ egalitatea unităţilor + aditivitate	Măsuri ale lungimii; perioade de timp

Scale nominale

La cel mai simplu nivel al acestei clasificări, Stevens a plasat ceea ce el a denumit scale *nominale*. Cuvântul *nominal* provine din rădăcina latină *nomen*, însemnând *nume*.

Prin urmare, în astfel de scale, numerele sunt utilizate doar ca etichete pentru a identifica un individ sau o clasă. Utilizarea nominală a numerelor pentru a eticheta indivizi este exemplificată de numerele de securitate socială (SS#) care identifică cei mai mulți oameni din Statele Unite; aceste numere sunt utile pentru că fiecare este atribuit unei singure persoane și, prin urmare, poate servi la identificarea persoanelor într-un mod mai specific decât prin nume și prenume, care pot fi comune mai multor oameni. Numerele pot fi utilizate și pentru a eticheta *date categoriale* care sunt date legate de variabile cum ar fi genul, afilierea politică, culoarea și așa mai departe — adică date care provin din distribuirea oamenilor, obiectelor sau evenimentelor în anumite categorii sau clase. De exemplu, atunci când introduc datele demografice în computer pentru analiză, cercetătorii creează de obicei o scală nominală care utilizează numere pentru a indica nivelele unei variabile categoriale. De exemplu, numărul 1 (unu) poate fi atribuit tuturor femeilor și 2 (doi) tuturor bărbaților. Singura cerință pentru această utilizare a numerelor este ca toți membrii unui set desemnat de un număr dat să fie egali cu privire la categoria căreia i se atribuie acel număr. Bineînțeles, deși numerele utilizate în scalele nominale pot fi adunate, scăzute, înmulțite sau împărțite, rezultatele unor astfel de operații sunt lipsite de sens. Atunci când utilizăm numere pentru a identifica anumite categorii, cum ar fi corect/greșit sau diagnostice psihiatrice, singura proprietate a unor astfel de numere este *identitatea*; aceasta înseamnă că tuturor

membrilor unei categorii trebuie să li se atribuie același număr și că nu pot exista două categorii care să aibă în comun același număr. Singura operație aritmetică permisă este numărarea frecvențelor din fiecare categorie. Bineînțeles, aceste frecvențe pot fi utilizate mai departe prin calcularea proporțiilor și prin realizarea unor analize ulterioare pe baza lor.

Scale ordinale

Numerele utilizate în scale *ordinale* transmit un înțeles ceva mai mare decât cele din scalele nominale, respectiv unul semnificativ. În aceste scale, pe lângă identitate, există proprietatea de *ordine a rangului*, ceea ce înseamnă că elementele dintr-un set pot fi grupate în serii — de la cel mai scăzut la cel mai înalt sau viceversa —, pot fi ordonate pe baza unei singure variabile, cum ar fi data nașterii sau nivelul performanței academice în cadrul unei clase date. Deși numerele de ordine a rangului transmit un înțeles precis în termenii poziției, ei nu poartă nici o informație cu privire la distanța dintre poziții. Astfel, studenții dintr-o clasă pot fi ordonați după rang în termenii performanței lor, dar această ordonare nu va reflecta mărimea diferenței dintre ei, care ar putea fi mare sau mică. În mod similar, în orice organizație ierarhică, să spunem U.S. Navy, rangurile (ex.: caporal, locotenent, comandant, căpitan, amiral) denotă diferite poziții, de la cea mai joasă la cea mai înaltă, dar diferențele dintre ele în termenii realizărilor sau prestigiului nu sunt aceleași. Dacă acestor ranguri li se atribuie numere, cum ar fi 1, 3, 7, 14 și 35, ordinea va fi menținută, dar nu se adăugă o semnificație suplimentară.

În testarea psihologică, utilizarea numerelor ordinale pentru a transmite rezultatele testării este generalizată. Scorurile la teste,

ordonate după rang, sunt prezentate ca *scoruri de rang percentil (RP)* care nu trebuie confundate cu scorurile procentuale familiare utilizate pe scară largă în notarea din învățământ.

Scorurile de rang percentil reprezintă pur și simplu o mulțime de numere ordinale ale unei scale de 100, astfel încât rangul indică procentul de indivizi dintr-un grup care intră la un anumit nivel sau sub un anumit nivel de performanță. De exemplu, un scor 70 al rangului percentil indică un nivel al performanței care este egal sau depășește pe cel atins de 70 % dintre oamenii din grupul în cauză. Scorurile de rang percentil, denumite deseori, pur și simplu, *percentile*, sunt vehiculul principal prin care utilizatorii testului transmit informațiile normative obținute din teste și, ca urmare, ele vor fi discutate din nou, într-o măsură mai largă, în următorul capitol.

Datele numerice din scalele ordinale pot fi manipulate statistic în același mod ca datele nominale. Suplimentar, există câteva tehnici statistice, cum ar fi coeficientul de corelație ρ — Spearman (r_S) pentru diferențe ale rangului, care sunt adecvate în mod specific utilizării în cazul unor date la nivel ordinal.

Scale de interval

În cazul scalelor de interval, numite și scale *cu unități egale*, numerele dobândesc încă o proprietate importantă. În aceste scale, diferențele dintre oricare două numere consecutive reflectă o diferență empirică sau demonstrabilă egală între obiectele sau evenimentele pe care numerele le reprezintă. Un exemplu este utilizarea zilelor pentru a marca trecerea timpului în calendar. O zi constă din 24 de ore, fiecare oră are 60 de minute, fiecare minut are 60 de secunde; dacă două date se află la o distanță de

12 zile, distanța dintre ele este exact de trei ori mai mare decât între două date care se află la 4 zile una de alta. Totuși, observați că un calendar exprimat în luni nu este o scală cu unități egale, deoarece unele luni sunt mai lungi decât altele. Mai mult, calendarul exprimă în mod tipic o caracteristică a scalelor de interval care limitează semnificația numerelor utilizate în ele, și anume că nu există niciun punct zero adevărat. În cazul calendarului, nu există un acord cu privire la punctul de plecare pentru începutul timpului. Diferite culturi au inventat puncte de început arbitrare, cum ar fi anul în care se presupune că s-ar fi născut Hristos, pentru a marca trecerea anilor. De exemplu, mult anticipata sosire a noului mileniu la sfârșitul anului 2000 al Erei Creștine sau erei noastre a avut loc în anul 5761 al calendarului evreiesc și în anul 4699 al calendarului chinezesc, ambele începând cu mulți ani înainte de începutul erei noastre.

În scalele de interval, distanțele dintre numere sunt semnificative. Astfel, putem aplica acestor numere cele mai multe operații aritmetice și putem obține rezultate care au sens. Cu toate acestea, datorită caracterului arbitrar al punctului de zero, numerele dintr-o scală de interval nu pot fi interpretate în termeni de raport.

Scale de raport

În cadrul scalelor de *raport* numerele dobândesc proprietatea de *aditivitate*, ceea ce înseamnă că ele pot fi adunate — ca și scăzute, înmulțite și împărțite — iar rezultatul poate fi exprimat ca raport, toate operațiile având rezultate pline de sens. Scalele de raport au un punct de zero absolut sau adevărat care stă pentru „nimic" din ceea ce este măsurat. În fizică, utilizarea acestui tip

de scală de măsură este obișnuită; timpul, distanțele, masa și volumele pot fi exprimate ca rapoarte în modalități semnificative și consistente logic. De exemplu, un obiect care cântărește 16 kilograme este de două ori mai greu decât unul care cântărește 8 kilograme (16/8=2), așa cum un obiect de 80 de kilograme este de două ori mai greu decât un obiect de 40 de kilograme (80/40 = 2). În plus, punctul de zero al scalei cântarului indică absența absolută a masei. În psihologie, scalele de raport sunt utilizate în principal atunci când măsurăm în termenii numărării frecvențelor sau a intervalelor de timp, ambele permițând posibilitatea de zero absolut.

Datele categoriale sau discrete pot fi măsurate — sau înregistrate — doar cu ajutorul scalelor nominale sau cu scale ordinale dacă datele se află într-un anumit tip de secvență. Datele continue sau metrice pot fi măsurate cu scale de interval sau cu scale de raport dacă există un punct de zero absolut. În plus, datele continue pot fi convertite în clase sau categorii și manipulate cu scale nominale sau ordinale. De exemplu, putem împărți oamenii în numai trei categorii — înalți, medii și scunzi — prin stabilirea unor puncte arbitrare de separare în variabila continuă a înălțimii. Atunci când trecem de la scale nominale la scale de raport, trecem de la numere care poartă o informație mai mică la numere care poartă mai multă informație. Ca o consecință a acestui fapt, trecerea de la un nivel al măsurării la altul ne cere să fim atenți dacă informația pe care numerele o presupun se păstrează în urma transformărilor sau manipulărilor pe care le aplicăm.

Nu uitați!

- În măsurare trebuie să existe o legătură demonstrabilă între numerele aplicate obiectelor, evenimentelor sau oamenilor și realitatea pe care numerele o reprezintă.
- Atunci când regulile utilizate pentru a crea această legătură nu sunt înțelese, rezultatele măsurării pot fi cu ușurință interpretate greșit.
- Pe măsură ce trecem de la un nivel de măsură la altul, trebuie să fim atenți dacă informația pe care numerele o conțin se păstrează în urma transformărilor sau manipulărilor pe care le aplicăm.
- Scorurile sunt numere cu semnificație specifică. În afara limitelor între care scorurile au semnificație, este probabil să se realizeze inferențe inexacte pornind de la aceste scoruri.

De ce este relevantă semnificația numerelor pentru testarea psihologică?

Deși nu este universal acceptat, sistemul lui Stevens de clasificare a scalelor de măsură ajută la menținerea *relativității* semnificației numerelor într-o perspectivă adecvată. Rezultatele celor mai multe teste psihologice sunt exprimate sub forma scorurilor, care sunt numere cu o semnificație specifică. Dacă limitele semnificației scorurilor nu sunt înțelese, este posibil ca pe baza acestor scoruri să fie realizate inferențe neadecvate. Din nefericire, acesta este un caz mult prea frecvent, așa cum se poate vedea din următoarele exemple.

Exemplul 1. Limite specifice ale scalelor ordinale. Așa cum am menționat anterior, multe scoruri sunt prezentate sub forma rangurilor percentile, care sunt numere la nivel ordinal care nu presupun egalitatea unităților. Dacă două scoruri sunt separate de 4 unități de rang percentil — ex. percentila 45 și 50 — diferența dintre ele și ce reprezintă această diferență în termenii a ceea ce este măsurat nu poate fi echivalată cu diferența ce separă oricare alte scoruri care se află la o distanță de 5 percentile — de exemplu, percentila 90 și 95. Într-o distribuție a scorurilor care aproximează

curba normală discutată mai târziu în acest capitol și reprezentată grafic în Figura 2.2, majoritatea scorurilor la test se grupează în jurul centrului distribuției. Aceasta înseamnă că, în astfel de distribuții, diferențele dintre scorurile de rang percentil sunt întotdeauna mai mari la extreme sau la cozile distribuției decât în mijloc.

Exemplul 2: Problema raportului IQ. Coeficienții de inteligență construiți inițial pentru a fi utilizați cu Scala de inteligență Stanford-Binet (S-B) au fost *raporturi IQ*. Adică, ei au fost *coeficienți* reali, obținuți prin împărțirea scorurilor vârstei mentale (VM) pe care un copil le-a obținut la testul S-B la vârsta cronologică (VC) a copilului și înmulțirea rezultatului cu 100 pentru a elimina zecimalele. Ideea era că un copil aflat la medie va avea vârste mentale și cronologice similare și un IQ de aproximativ 100. Copiii care funcționează sub medie vor avea vârste mentale mai scăzute decât vârstele cronologice, și IQ-ul sub 100, în timp ce copiii care funcționează peste medie vor avea o vârstă mentală mai mare decât vârsta cronologică și IQ-ul peste 100. Această idee a funcționat bine pentru copii de vârsta școlii primare și gimnaziale, perioadă în care există tendința unui ritm stabil al creșterii intelectuale de la un an la altul. Cu toate acestea, raportul VM/VC pur și simplu nu funcționează pentru adolescenți și adulți, datorită faptului că dezvoltarea lor intelectuală este mult mai puțin uniformă — și schimbările sunt deseori imperceptibile de la un an la altul. Faptul că vârsta cronologică maximă utilizată în calcularea raportului IQ în scala originală S-B a fost 16 ani, indiferent de vârsta reală a persoanei testate, a creat probleme suplimentare de interpretare. Mai mult, scalele vârstei mentale și ale vârstei cronologice nu se află la același nivel de măsură. Vârsta mentală, așa cum a fost evaluată prin intermediul primelor teste de inteligență, reprezintă

în mod fundamental un nivel ordinal de măsură, în timp ce vârsta cronologică poate fi măsurată pe o scală de raport. Din aceste

Info 2.2

Exemple de calcul al raportului IQ și problemele care apar în această situație

Subiect	Vârstă mentală	Vârstă cronologică (VM)	Diferență (VM-VC) (VC)	Raport IQ[a]
Ally	6 ani	5 ani	1 an	6/5 x 100 = 120
Ben	12 ani	10 ani	2 ani	12/10 x 100 = 120
Carol	18 ani	15 ani	3 ani	18/15 x 100 = 120

[a] În calculul concret al raporturilor IQ, atât vârsta mentală, cât și cea cronologică au fost exprimate în luni, nu în ani.

Problema 1: Scorurile vârstei mentale necesare pentru a obține un IQ dat cresc continuu pentru fiecare vârstă cronologică succesivă, astfel încât rapoartele IQ la diferite vârste cronologice nu sunt echivalente.

Problema 2: În timp ce vârsta cronologică crește constant, nu se întâmplă același lucru cu vârsta mentală. Din moment ce cea mai ridicată vârstă mentală ce poate fi atinsă la un test de inteligență dat nu poate fi nelimitată, chiar și atunci când se pune o limită vârstei cronologice maxime utilizate pentru a calcula IQ — așa cum s-a întâmplat mult timp în cazul scalei S-B —, IQ-urile pe care le pot obține cei mai mulți adulți sunt limitate artificial comparativ cu cele obținute de copii și adolescenți.

Soluție: Datorită acestor probleme și a altora legate de raportul IQ și de conceptul de vârstă mentală, utilizarea raportului IQ a fost abandonată. Termenul *IQ* este utilizat acum pentru un scor care nu este un raport IQ și nu este nici măcar un coeficient. Acest scor, cunoscut ca *deviație IQ*, a fost introdus de David Wechsler și este explicat în Capitolul 3.

motive, împărțirea unui număr la altul pentru a obține un coeficient nu conduce la rezultate cu sens, consistente logic.

Info 2.2 prezintă exemple numerice care evidențiază unele dintre aceste probleme care au făcut ca raportul IQ să fie abandonat.

Ce putem conchide cu privire la semnificația numerelor în măsurătorile psihologice?

În psihologie este esențial să păstrăm în minte faptul că cea mai mare parte a scalelor noastre de măsură sunt de natură ordinală. Egalitatea unităților este aproximată de scalele utilizate pentru multe tipuri de scoruri ale testelor, dar o astfel de egalitate nu este niciodată la fel de permanentă și completă ca în fizică, deoarece unitățile sunt legate de performanțele eșantioanelor din care sunt obținute. Utilizarea scalelor de raport în psihologie este limitată la măsurarea frecvențelor, a timpului de reacție sau a variabilelor care pot fi exprimate cu sens prin unități fizice. De exemplu, dacă am utilizat rezultatele pe oră de pe linia de asamblare 1, putem spune că Muncitorul A, care produce 15 unități pe oră, este de trei ori mai rapid decât Muncitorul B, care produce doar 5 unități pe oră. Totuși, observați că nu putem spune că Muncitorul A este de trei ori mai bun decât Muncitorul B, deoarece viteza nu este singurul indicator al performanței în muncă, chiar și în cazul operării la o linie de asamblare. Nivelul global al performanței este o variabilă mai complexă care, cel mai probabil, poate fi evaluată doar cu o scală calitativă, ordinală.

Tipuri de statistici

Deoarece utilizarea numerelor pentru a reprezenta obiecte și evenimente este atât de răspândită în testarea psihologică, domeniul implică aplicații substanțiale ale *statisticii*, o ramură a matematicii dedicată organizării, descrierii, sintetizării, analizării și manipulării datelor numerice. Numerele și graficele utilizate pentru a descrie, condensa sau reprezenta datele aparțin domeniului *statisticii descriptive*. Pe de altă parte, atunci când datele sunt utilizate pentru a estima valorile populației pe baza valorilor eșantionului sau pentru a testa ipoteze, este aplicată *statistica inferențială* — un set mai amplu de proceduri bazat pe teoria probabilităților. Din fericire, deși atât statistica inferențială, cât și statististica descriptivă sunt utilizate extensiv în construirea testelor, cele mai multe dintre aspectele cantitative ale interpretării scorurilor la un test necesită doar o bună înțelegere a statisticii descriptive și a unui număr relativ redus de tehnici de tip inferențial. Mai mult, chiar dacă este de dorit un fundament în matematici superioare pentru a înțelege pe deplin statistica implicată în testarea psihologică, este posibil să o înțelegem la un nivel bazal cu o bună doză de logică și cunoștințe de matematică relativ limitate.

Cuvintele *statistică* și *statistici* sunt utilizate și pentru a face referiri la măsurători derivate din datele eșantionului —

Nu uitați!

Cele două sensuri ale *statisticii*

1. Studierea și aplicarea metodelor de organizare, descriere, rezumare, analiză și alte manipulări ale datelor.
2. Numere (ex. medii, coeficienți de corelație) care descriu caracteristicile variabilelor sau ale mulțimilor de date obținute din eșantioane, în opoziție cu cele obținute din populații, numite *parametri*.

ca opuse celor obținute de la populații, care sunt denumite *parametri*. Mediile, abaterile standard, coeficienții de corelație și alte astfel de numere calculate pornind de la datele eșantionului sunt toate statistici obținute pentru a estima ceea ce este de real interes, și anume, parametrii populației respective. *Parametrii* sunt numere matematice exacte (sau constante, cum ar fi π) care nu pot fi obținute de obicei decât dacă populația este atât de stabilă și circumscrisă, încât toți membrii săi pot fi descriși, cum ar fi toți membrii unei clase de liceu într-un semestru dat. De fapt, unul dintre scopurile principale ale statisticii inferențiale este de a estima parametrii populației pe baza datelor eșantionului și a teoriei probabilităților.

Statistica descriptivă

Datele brute sunt greoaie. De obicei, ele constau dintr-o grămadă de numere care nu transmit niciun sens, nici măcar după o examinare atentă, cum ar fi cele 60 de numere prezentate în Tabelul 2.2. Aceste numere reprezintă scorurile obținute de 60 de elevi de liceu la primul test dat la un curs de testare psihologică; testul constă din 50 de itemi cu alegere multiplă. O simplă privire peste numerele din tabel aduce în atenție anumite informații, cum ar fi faptul că cele mai multe scoruri se situează undeva între 30 și 50. Cu ajutorul statisticii descriptive putem sintetiza datele, astfel încât ele devin mai ușor de înțeles. O cale de a rezuma datele este de a le reprezenta grafic; o altă cale este de a le condensa în statistici care prezintă numeric informația conținută într-un set de date.

Tabelul 2.2 Date brute: 60 de scoruri la test

41	50	39	40	40	31	42	29	37	36
35	45	44	49	38	34	35	32	41	41
39	47	30	45	43	47	35	46	42	41
34	37	38	40	39	39	36	32	48	39
33	42	44	48	47	40	33	46	46	40
44	37	45	43	39	42	37	45	43	38

Distribuția frecvențelor

Înainte de a aplica orice formulă statistică, este întotdeauna o idee bună ca datele brute să fie organizate practic, astfel încât să poată fi examinate. În mod normal, acest fapt se realizează prin intermediul *distribuției frecvențelor*. Tabelul 2.3 prezintă o distribuție a scorurilor testului din Tabelul 2.2, prezentând numărul de câte ori sau frecvența cu care a apărut fiecare scor și procentul apariției sale. Coloana Procent cumulat prezintă suma numerelor din coloana Procent de la scorurile cele mai scăzute la cele mai ridicate. Această ultimă mulțime de figuri ne permite să vedem procentele pentru cele 60 de cazuri care se situează la nivelul fiecărui scor sau sub acest scor și astfel pot fi citite cu ușurință drept scoruri de rang percentil pentru indivizii care au obținut fiecare scor.

Atunci când intervalul scorurilor este foarte mare, *distribuțiile frecvențelor grupate* ajută la organizarea scorurilor într-o formă mai compactă. În aceste distribuții, scorurile sunt grupate în intervale cu o mărime convenabilă pentru a cuprinde datele, iar frecvențele sunt prezentate pentru fiecare interval, nu pentru fiecare scor. Bineînțeles, ceea ce se câștigă prin caracterul compact se pierde în termeni de detaliu al informației.

Tabelul 2.3 Distribuția de frecvențe a 60 de scoruri la test

Scoruri	Frecvența (*f*)	Procent (*P*)	Procent cumulat (*PC*)[a]
50	1	1,7	100,0
49	1	1,7	98,3
48	2	3,3	96,7
47	3	5,0	93,3
46	3	5,0	88,3
45	4	6,7	83,3
44	3	5,0	76,7
43	3	5,0	71,7
42	4	6,7	66,7
41	4	6,7	60,0
40	5	8,3	53,3
39	6	10,0	45,0
38	3	5,0	35,0
37	4	6,7	30,0
36	2	3,3	23,3
35	3	5,0	20,0
34	2	3,3	15,0
33	2	3,3	11,7
32	2	3,3	8,3
31	1	1,7	5,0
30	1	1,7	3,3
29	1	1,7	1,7

[a]Rotunjit la cea mai apropiată zecimală

Grafice

Din momentul în care datele au fost organizate într-o distribuție a frecvențelor, ele pot fi transpuse în una sau mai multe forme grafice, cum ar fi diagrame circulare sau cu coloane (pentru date discrete sau categoriale) și histograme sau poligoane de

frecvență (pentru date continue sau metrice). Datele din Tabelul 2.3 sunt prezentate grafic sub forma unui poligon de frecvențe în Figura 2.1. De obicei, axa orizontală (numită și *abscisă*, *linia de bază* sau *axa X*) este folosită pentru a reprezenta intervalul de valori ale variabilei discutate și axa verticală (numită și *ordonată* sau *axa Y*) pentru a ilustra frecvența cu care fiecare dintre valori apare în distribuție. Regulile și procedeele de transformare a distribuțiilor de frecvențe în diferite tipuri de grafice sunt prezentate în cele mai multe cursuri de fundamente de statistică (vezi, Kirk, 1999).

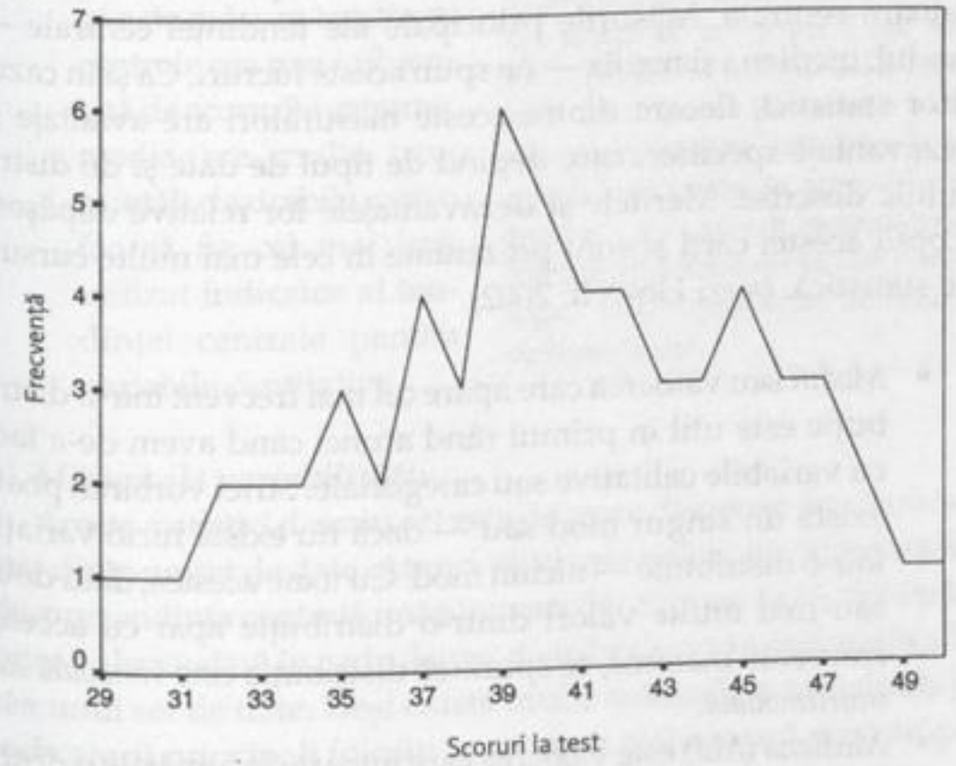

Figura 2.1 Poligonul de frecvențe pentru scorurile la test din Tabelul 2.3 (n = 60)

Pe lângă faptul că ne ajută să vizualizăm datele prin grafice, statisticile descriptive ne oferă și instrumentele care ne permit sintetizarea numerică a proprietăților datelor. Aceste instrumente descriu tendința centrală și variabilitatea datelor numerice.

Măsuri ale tendinței centrale

Unul dintre primele lucruri pe care cineva vrea să le cunoască atunci când inspectează un set de date este localizarea celei mai mari aglomerări de date, și datele cele mai reprezentative sau cu valoare centrală. Măsurile principale ale tendinței centrale — modul, mediana și media — ne spun aceste lucruri. Ca și în cazul altor statistici, fiecare dintre aceste măsurători are avantaje și dezavantaje specifice, care depind de tipul de date și de distribuțiile descrise. Meritele și dezavantajele lor relative depășesc scopul acestei cărți și sunt prezentate în cele mai multe cursuri de statistică. (vezi Howell, 2002).

- *Modul* sau valoarea care apare cel mai frecvent într-o distribuție este util în primul rând atunci când avem de-a face cu variabile calitative sau categoriale. Strict vorbind, poate exista un singur mod sau — dacă nu există nicio variație într-o distribuție — niciun mod. Cu toate acestea, dacă două sau mai multe valori dintr-o distribuție apar cu aceeași frecvență maximă, se spune că distribuția este *bimodală* sau *multimodală*.
- *Mediana* (*Md*) este valoarea care împarte în jumătate o distribuție care a fost ordonată după mărime. Dacă numărul valorilor (*n*) distribuției este impar, mediana este pur și simplu valoarea din mijloc; dacă *n* este par, mediana se află în punctul de la jumătatea distanței dintre cele două valori de mijloc.
- *Media* sau media aritmetică (μ pentru media populației și M pentru media eșantionului) este obținută prin însumarea tuturor valorilor dintr-o distribuție și împărțirea totalului la numărul cazurilor din distribuție. Astfel, valoarea sa concretă poate să fie sau să nu fie prezentă în setul de date. În ciuda acestui fapt și a faptului că este măsura tendinței centrale cea mai influențată de scorurile extreme, media are multe proprietăți dezirabile care o fac să fie cel mai larg utilizat indicator al tendinței centrale pentru variabile cantitative.

> **ATENȚIE!**
>
> În paginile care urmează veți întâlni câteva formule statistice. Dacă sunteți tentați să treceți peste ele, NU FACEȚI ASTA. Amintiți-vă, aceasta este o carte despre *Esențialul în testarea psihologică*. Singurele formule pe care le veți întâlni în această carte sunt cele care transmit concepte *esențiale* pentru înțelegerea testării psihologice și a semnificației scorurilor obținute la test.

Măsuri ale variabilității

Aceste statistici descriu cât este de mare dispersia sau *împrăștierea* dintr-un set de date. Atunci când este adăugată informațiilor despre tendința centrală, măsura variabilității ne ajută să plasăm orice valoare dată în cadrul unei distribuții și să sporim descrierea unui set de date. Deși există multe măsuri ale variabilității, indicatorii principali folosiți în testarea psihologică sunt amplitudinea, amplitudinea semi-intercuartilică, dispersia și abaterea standard.

- *Amplitudinea* este distanța dintre două puncte extreme — valoarea cea mai mare și cea mai mică — dintr-o distribuție. Deși amplitudinea este calculată cu ușurință, este o măsură foarte instabilă, căci se poate schimba drastic datorită prezenței unuia sau a două scoruri extreme.
- *Amplitudinea semi-intercuartilică* reprezintă o jumătate din *amplitudinea intercuartilică* (AIQ) care, la rândul său, reprezintă distanța dintre punctele care demarchează limitele superioare ale primului și celui de-al treilea sfert al unei distribuții. Primul punct de cuartilă (Q_1) sau percentila 25, marchează partea superioară a sfertului inferior (cuartilă) al distribuției. Al treilea punct de cuartilă (Q_3) sau percentila 75 se află la capătul superior al celui de-al treilea sfert al distribuției și marchează începutul cuartilei superioare. Amplitudinea intercuartilică este intervalul dintre Q_1 și Q_3; prin urmare, el cuprinde 50% dintr-o distribuție, partea de mijloc a acesteia. În exemplul prezentat în Tabelul 2.3, percentila 25 este la scorul 37, iar percentila 75 se află la scorul 44. Amplitudinea intercuartilică este $44 - 37 = 7$ și amplitudinea semi-intercuartilică este $7 \div 2 = 3{,}5$. Observați că, în timp ce 53% din scoruri intră într-un interval redus de 8 puncte, celelalte 47% din scoruri sunt răspândite în întregul interval rămas de 14 puncte de scor.
- *Dispersia* reprezintă suma pătratelor diferențelor sau a abaterilor dintre fiecare valoare (X) dintr-o distribuție și media acelei distribuții (M), împărțită la N. Mai succint, dispersia este media *sumei pătratice* (*SP*). Suma pătratică este o abreviere pentru suma pătratelor valorilor abaterii sau a scorurilor abaterii, $\Sigma(X - M)^2$. Scorurile abaterii trebuie să fie ridicate la puterea a doua înainte de a fi adunate pentru a

elimina numerele negative. Dacă aceste numere nu ar fi ridicate la pătrat, scorurile pozitive și negative ale abaterilor din jurul mediei s-ar anula unele pe altele și suma lor ar fi zero. Suma pătratică reprezintă mărimea totală a variabilității într-o distribuție de scoruri și dispersia (SP/N) reprezintă variabilitatea medie. Totuși, datorită ridicării la pătrat a scorurilor abaterii, dispersia nu este exprimată în aceleași unități ca distribuția inițială.

- *Abaterea standard* este rădăcina pătrată a dispersiei. Alături de dispersie, ea oferă o singură valoare reprezentativă pentru diferențele individuale sau abaterile dintr-un set de date — calculate dintr-un punct de referință comun, și anume media. Abaterea standard este o estimare a variabilității medii a unui set de scoruri, exprimată în aceleași unități ca și scorurile. Este măsura cea mai tipică a variabilității utilizată în testare și în multe alte scopuri și este utilă în diferite manipulări statistice ale datelor.

Info 2.3 enumeră câteva dintre simbolurile fundamentale care vor fi utilizate în această carte, alături de formulele pentru medie, amplitudine intercuartilică, abaterea standard și dispersie. Măsurile tendinței centrale și ale variabilității pentru cele 60 de scoruri obținute la un test din Tabelul 2.3 sunt prezentate în Tabelul 2.4. Deși informațiile detaliate despre cele 60 de scoruri sunt pierdute, statisticile din Tabelul 2.4 descriu concis locul în care scorurile se grupează și gradul mediu al dispersiei în setul de date.

Importanța variabilității

Deși poate fi adevărat că varietatea este „sarea și piperul" vieții, ea este chiar partea principală a testării. Testarea psihologică este determinată de variabilitatea indivizilor. Fără diferențe indivi-

Info 2.3

Notații și formule de bază

X = Un singur punct al datelor sau o singură valoare dintr-o distribuție; în testarea psihologică, aproape întotdeauna X stă pentru un scor brut.

Σ = Suma din...

n = mărimea eșantionului, adică numărul total de cazuri dintr-o distribuție; în testarea psihologică, n reprezintă aproape întotdeauna numărul de persoane sau numărul de scoruri.

N = mărimea populației

M_X sau $\bar{X}$ = media lui X = $\frac{\Sigma X}{n}$

μ = media populației

Mdn = Mediana = percentila 50

$Q1$ = primul punct de cuartilă = percentila 25

$Q3$ = al treilea punct de cuartilă = percentila 75

$Q1 - Q3$ = amplitudine intercuartilică (AIQ).

$AIQ \div 2$ = amplitudine semi-intercuartilică

s^2 = dispersia eșantionului = $\frac{\Sigma(X - M)^2}{n-1}$

σ^2 = dispersia populației = $\frac{\Sigma(X - \mu)^2}{N}$

s sau SD = abaterea standard a eșantionului = $\sqrt{s^2}$

σ = abaterea standard a populației = $\sqrt{\sigma^2}$

Tabelul 2.4 Statistica descriptivă pentru cele 60 de scoruri ale testului din Tabelul 2.3

Măsuri ale tendinței centrale		Măsuri ale variabilității	
Media	= 40,13	Interval = 50 – 29	= 21
Mediana	= 40,00	Dispersie	= 25,745
Modul	= 39	Abaterea standard	= 5,074
Q_1 sau percentila 25	= 37	Amplitudine intercuartilică = $= Q_3 - Q_1 = 44 - 37$	= 7
Q_3 sau percentila 75	= 44	Amplitudine semi-intercuartilică = $= 7 \div 2$	= 3,5

duale nu ar exista nicio variabilitate și testele ar fi inutile pentru a ne ajuta să tragem concluzii sau să luăm decizii cu privire la oameni. Dacă toate celelalte aspecte rămân neschimbate, cu cât este mai mare gradul de variabilitate existent între indivizi cu privire la o caracteristică pe care încercăm să o evaluăm, cu atât mai precis putem face distincțiile necesare între ei. Cunoașterea formei distribuțiilor scorurilor, ca și a tendințelor centrale și variabilității oferă o bază pentru o bună parte a esențialului despre interpretarea scorurilor, discutat în Capitolul 3.

Exersați

- Mergeți la Tabelul 2.3 și numărați câte scoruri se găsesc în intervalul ± 1 SD față de medie, adică între 40 ± 5.
- Se demonstrează că 41 din cele 60 de scoruri, aproximativ 2/3 din ele, se află între 35 și 45.
- Această proporție este tipică pentru distribuții care aproximează forma curbei normale cum este distribuția din Figura 2.1.

Modelul curbei normale

Definiție

Curba normală, cunoscută și drept *curba „clopot"*, este o distribuție care, în anumite moduri, este similară celei din figura 2.1. Abscisa, echivalentă cu axa x a distribuției din Figura 2.1, indică unitățile abaterii standard (σ); axa sa verticală, sau ordonata, de obicei nu e nevoie să fie reprezentată, deoarece curba normală nu este o distribuție de frecvență a datelor, ci un model matematic al unei distribuții ideale sau teoretice. Înălțimea la care se ridică curba în fiecare punct al axei orizontale este determinată de o formulă matematică ce descrie relația specifică din cadrul modelului și stabilește forma exactă și proporțiile curbei. Ca orice model ideal, curba normală nu există; ea are la bază teoria probabilităților. Din fericire, pentru scopurile noastre, putem înțelege aspectele fundamentale ale curbei normale fără să știm multe despre bazele sale matematice.

Deși modelul curbei normale este un ideal, deseori este aproximat de distribuții ale datelor reale, cum ar fi datele din Tabelul 2.3, reprezentate în Figura 2.1. Similaritatea dintre model și distribuțiile mai multor variabile din lumea naturală a făcut ca acesta să fie util în statisticile descriptive. Mai important însă este faptul că multe evenimente întâmplătoare, dacă se repetă de suficiente ori, generează distribuții care aproximează curba normală. Tocmai datorită acestei legături cu teoria probabilităților, curba normală are un rol important în statisticile inferențiale. Așa cum vom vedea, utilitatea modelului curbei normale derivă din proprietățile sale către care ne vom îndrepta atenția acum.

Multe dintre proprietățile modelului curbei normale sunt evidente în mod clar după o examinare vizuală (vezi Figura 2.2). De exemplu, se poate observa că distribuția normală are următoarele proprietăți:

- Are *formă de clopot*, așa cum indică și numele ei
- Este *bilateral simetrică*, ceea ce înseamnă că cele două părți ale sale sunt identice (dacă împărțim curba în două, fiecare jumătate conține 50% din aria de sub curbă).
- Are cozi care se apropie de abscisă, dar niciodată nu o ating, și astfel *limitele sale se extind la ± infinit* (±∞), o proprietate care subliniază natura teoretică și matematică a curbei.

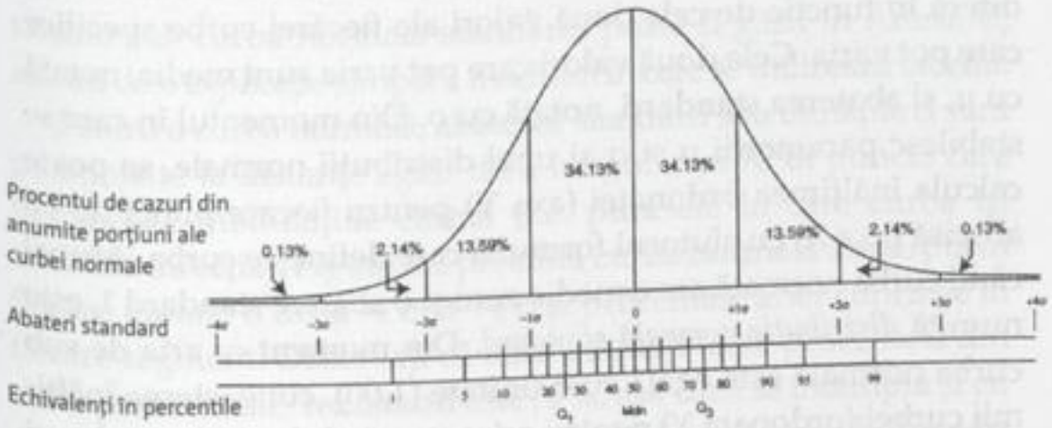

Figura 2.2 Curba normală cu procentele de cazuri din fiecare segment de unități σ de la –4 la +4, procente cumulate și echivalenții în percentile.

- Este *unimodală*; adică are un singur punct de frecvență maximă sau înălțime maximă.
- Are o *medie, mediană* și *un mod* care coincid în centrul distribuției pentru că punctul în care curba este într-un echilibru perfect, care este media, este, de asemenea, și punctul care împarte curba în două părți egale, care reprezintă mediana și este și cea mai frecventă valoare, care reprezintă modul.

Pe lângă aceste proprietăți, curba normală are și alte caracteristici, mai puțin evidente, care sunt legate de formula funcției sale matematice. Această formulă — care nu este esențială — poate fi găsită în cele mai multe manuale de statistică și pe unele dintre site-urile Web destinate curbei normale menționate în Info 2.4. Formula presupune două elemente constante (σ și e) și două valori care pot varia. Fiecare curbă normală particulară este doar un caz al unei familii de distribuții ale curbei normale care diferă în funcție de cele două valori ale fiecărei curbe specifice care pot varia. Cele două valori care pot varia sunt media, notată cu μ, și abaterea standard, notată cu σ. Din momentul în care se stabilesc parametrii μ și σ ai unei distribuții normale, se poate calcula înălțimea ordonatei (axa Y) pentru fiecare punct de pe abscisă (axa X) cu ajutorul formulei care definește curba. Atunci când curba normală are o medie zero și o abatere standard 1, este numită *distribuție normală standard*. Din moment ce aria de sub curba normală este egală cu o unitate (1,00), cunoașterea înălțimii curbei (ordonata Y) pentru orice punct al abscisei sau al axei X ne permite să calculăm proporția (*p*) sau procentul (*p* x 100) din aria de sub curbă care se află peste și sub orice valoare X, precum și aria dintre oricare două valori X. Tabelul statistic care rezultă în urma acestor calcule, în care sunt prezentate ariile

Info 2.4

Site-uri Web despre Curba normală

Acestea sunt doar trei dintre numeroasele site-uri care pot fi găsite introducând „the normal curve" într-un motor de căutare online bun:

- http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html

Acest site are o demonstrație simplă și atractivă vizual a procesului care duce la curba normală.

- http://stat-www.berkeley.edu/~stark/Java/NormHiLite.htm

Această pagină Web are un instrument interactiv care îți permite să selectezi orice segment al curbei normale și să vezi imediat ce procent al ariei se află în segmentul selectat.

- http://www.psychstat.smsu.edu/introbook/sbk11.htm

Acesta este unul dintre cele câteva site-uri care explică aspectele fundamentale ale curbei normale în mod clar și succint.

situate sub curba normală standard, poate fi găsit în Anexa C, alături de o explicație simplă a modului în care se utilizează tabelul.

Pentru o curbă normală, abaterea standard sau unitățile σ sunt poziționate la distanțe egale de-a lungul axei X, în puncte care marchează modulațiile curbei (i.e. punctele în care curba își schimbă direcția). Figura 2.2 prezintă curba normală împărțită în fiecare unitate σ de la -4 σ la +4 σ și procentele ariei cuprinse în fiecare segment. Observați că dacă adunați toate procentele din ariile peste medie, rezultatul este 50%, așa cum se întâmplă și cu suma tuturor ariilor aflate sub medie. Mai mult, aria dintre +1σ și -1σ este de 68,26% (34,13% x 2) — aproximativ 2/3 din curbă — și aria dintre + 2σ și -2σ este 95,44%, aproape întreaga curbă. Cunoașterea acestor aspecte fundamentale ale curbei normale este extrem de utilă în statistică.

Utilizările modelului curbei normale

Utilizări descriptive

Deoarece proporţiile ariei aflate sub curba normală standard care se află peste sau sub orice punct al abscisei sau între două puncte ale abscisei sunt prestabilite — şi pot fi găsite cu uşurinţă în tabelele ariilor situate sub curba normală cum este cel din Anexa C — putem aplica fără dificultate aceste proporţii oricărei alte distribuţii care are o formă similară. În testare, această aplicaţie particulară a distribuţiei normale este utilizată în mod repetat pentru a genera scoruri standard, descrise în următorul capitol.

În anumite circumstanţe, chiar şi atunci când o distribuţie aproximează, dar nu corespunde exact curbei normale, putem utiliza proporţiile din modelul curbei normale pentru a normaliza scorurile. *Normalizarea* scorurilor presupune transformarea lor, astfel încât să aibă aceeaşi semnificaţie, în termenii poziţiei, ca şi atunci când ar proveni dintr-o distribuţie normal. Această procedură, care nu este atât de complicată pe cât pare, utilizează procentele cumulate calculate din distribuţia frecvenţelor (vezi Tabelul 2.3). Procedeul va fi discutat detaliat, însoţit de exemple, în capitolul următor.

Utilizări inferenţiale ale modelului curbei normale

În statistica inferenţială, modelul curbei normale este util pentru a) estimarea parametrilor populaţiei şi b) testarea ipotezelor referitoare la diferenţe. Aplicaţiile modelului curbei normale la estimarea parametrilor populaţiei şi la testarea ipotezelor utilizează două noţiuni interrelaţionate, şi anume distribuţia de eşantionare şi eroarea standard.

Tabelul 2.5 Eroarea standard a mediei pentru datele din Tabelele 2.3 şi 2.4

Media (M) = 40,13	Abaterea standard(s) = 5,074	Mărimea eşantionului (n) = 60

Eroarea standard a mediei $(SE_M) = \frac{s}{\sqrt{n}} = \frac{5,074}{\sqrt{60}} = \frac{5,074}{7,7459} = 0,655$

Distribuţiile de eşantionare sunt distribuţii ipotetice de valori, opuse celor reale, care se bazează pe presupunerea că dintr-o populaţie poate fi extras un număr infinit de eşantioane de o anumită dimensiune. Dacă s-ar face acest lucru şi dacă s-ar înregistra statistica pentru fiecare dintre aceste eşantioane, multe (dar nu toate) dintre distribuţiile rezultate ale statisticilor sau distribuţiile de eşantionare ar fi normale. Media fiecărei distribuţii de eşantionare ipotetice va fi egală cu parametrul populaţiei şi abaterea standard a distribuţiei de eşantionare va fi eroarea standard a statisticii în chestiune.

Eroarea standard (*SE*) a statisticii eşantionului obţinut este concepută ca abatere standard a distribuţiei de eşantionare care ar rezulta dacă am obţine aceeaşi statistică dintr-un număr mare de eşantioane de mărime egală extrase aleator. Poate fi calculată cu uşurinţă utilizând statistica eşantionului (vezi Tabelul 2.5). Odată ce am obţinut o statistică dată şi eroarea sa standard pentru

Nu uitaţi!

Anexa C conţine Tabelul ariilor situate sub curba normală, împreună cu o explicare a modului în care poate fi folosit. Aşa cum se întâmplă cu toate formulele din această carte, informaţia din Anexa C este prezentată doar datorită faptului că familiarizarea cu ea este o cerinţă *esenţială* pentru înţelegerea scorurilor obţinute la un test.

un eșantion, presupunerea că distribuția de eșantionare este normală ne permite să utilizăm ariile situate sub curba normală pentru a estima parametrii populației pe baza statisticii obținute.

Estimarea parametrilor populației

Un exemplu ipotetic

Pentru a estima un parametru al populației, cum ar fi înălțimea medie a femeilor adulte din Statele Unite, am putea obține un eșantion extras aleator alcătuit din 50 de femei adulte, una din fiecare stat. Să presupunem că înălțimea medie pentru acel eșantion, care ar fi o *estimare* a mediei populației, este de 64 inch și să presupunem că abaterea standard este de 4 inch. Dacă am repeta această procedură, extrăgând un număr infinit de eșantioane alcătuite din 50 de femei și înregistrând de fiecare dată mediile pentru aceste eșantioane, distribuția de eșantionare a mediilor care ar rezulta s-ar potrivi modelului curbei normale. Media acestei distribuții de eșantionare teoretice ar fi interpretată ca medie a populației (i.e. înălțimea medie a femeilor adulte din Statele Unite).

În mod evident, un astfel de parcurs al acțiunii nu este doar nepractic, dar este și imposibil. Ca urmare, utilizăm statistici inferențiale pentru a estima media populației. Obținem *eroarea standard a mediei* (SE_M) cu formula $s/\sqrt{n}$, unde s este abaterea standard pentru eșantion (4) iar n este numărul de cazuri din eșantion (50). În acest caz, împărțind 4 la rădăcina pătrată din 50 obținem 7,07, iar $SE_M = 0{,}565$.

Astfel, pe baza statisticii pentru eșantion, am putea spune că înălțimea medie a femeilor adulte din Statele Unite se află în intervalul mediei obținute de noi de 64 inch ±0,565 inch sau între

63,435 și 64,565 inch. Adăugând și scăzând $1SE_M$ din media eșantionului obținem un interval de încredere de 68% pentru media populației, deoarece 68% din aria de sub curba normală și află între ± 1σ (sau, în acest caz, ±1 SE_M). Dacă dorim să facem o afirmație cu un nivel mai mare de încredere, am putea alege un interval de încredere mai mare, selectând un număr mai mare de unități σ și înmulțindu-l cu SE_M. Așa cum observăm în Figura 2.2, segmentul dintre ±2σ cuprinde 95,44% din aria de sub curba normală; prin urmare, în exemplul nostru, am putea avea o încredere de 95,44% că se află înălțimea medie a femeilor adulte din Statele Unite în intervalul dintre 64 ±2 SE_M sau 64 ±2 (0,565 inch) = 64 ±1,13 inch care cuprinde valorile dintre 62,87 și 65,13 inch.

Exemplu cu datele din Tabelul 2.3: calculăm eroarea standard a mediei (SE_M) cu formula $s/\sqrt{n}$, unde s este abaterea standard

Nu uitați!

Două concepte esențiale ale statisticii inferențiale: distribuții de eșantionare și erori standard

- *Distribuțiile de eșantionare* sunt distribuții *teoretice* ale valorilor unei variabile sau ale unei statistici care ar rezulta dacă s-ar colecta și înregistra valorile (scoruri) sau statisticile (mediile, abaterile standard, coeficienții de corelație etc.) pentru un număr infinit de eșantioane de o mărime dată dintr-o anumită populație. Distribuțiile de eșantionare nu există cu adevărat; ele sunt construcții ipotetice care sunt utilizate pentru a estima probabilitatea valorilor sau statisticilor obținute prin intermediul *erorii standard*.
- *Erorile standard* sunt entități statistice care *pot* fi calculate prin diferite formule pe baza datelor eșantionului; ele oferă mijloacele prin care valorile obținute sau statisticile eșantionului sunt comparate cu distribuțiile lor teoretice de eșantionare. O eroare standard este abaterea standard estimată a distribuției de eșantionare a unei valori sau statistici obținute.

(5,074) și n este numărul de cazuri din eșantion (60), $SE_M = 0{,}655$ (vezi Tabelul 2.5). Dacă eșantionul de 60 de studenți a fost extras aleator din toți studenții cărora li s-a dat vreodată acel test, am putea presupune că există o probabilitate de aproximativ 68% ca media pentru populația *tuturor* studenților care au dat acel test să fie între media obținută de 40,13 ± 0,655 puncte, sau ±1 SE_M, sau undeva între 39,475 și 40,785. În mod similar, putem spune cu încredere de 95,44% — însemnând că șansele să greșim sunt mai mici de 5% — că intervalul dintre media de 40,13±2 SE_M, adică intervalul care se întinde de la 38,82 la 41,44 include media populației.

Semnificația erorilor standard

Erorile standard sunt extrem de importante în statisticile inferențiale. În ambele exemple pe care tocmai le-am prezentat, putem estima intervalele în care e probabil să poată fi găsiți parametrii populației pe baza premiselor că (a) media eșantionului obținută este cea mai bună estimare a mediei populației pe care o avem și (b) eroarea standard a mediei este echivalentă cu abaterea standard a distribuției de eșantionare ipotetice a mediilor, presupusă a fi normală. Presupuneri similare, alături de estimări ale ariilor situate sub curba normală și a altor distribuții teoretice care apar în manualele de statistică — cum ar fi distribuția *t*-Student — pot fi utilizate nu numai pentru a genera afirmații de probabilitate despre parametrii populației derivați din alte statistici ale eșantionului, dar și pentru a genera afirmații de probabilitate despre diferențele obținute între statisticile eșantionului.

Atunci când diferențele dintre mediile sau proporțiile eșantionului sunt testate pentru semnificația statistică, diferențele

obținute sunt împărțite la erorile standard ale acelor diferențe, calculate prin formule adecvate tipului specific de diferență care este testată. Raportul rezultat, numit *raport critic*, alături de distribuțiile adecvate pentru statisticile în chestiune, poate fi folosit pentru a stabili dacă diferența obținută ar fi putut rezulta întâmplător. Deși multe dintre tehnicile statisticilor inferențiale depășesc scopul acestei lucrări, vom reîntâlni erorile standard în legătură cu fidelitatea și validitatea scorurilor obținute la test, în capitolele 4 și 5. Info 2.5 rezumă motivele principale pentru care modelul curbei normale este atât de important în domeniul testării psihologice.

Distribuții ne-normale

Realizarea reprezentărilor grafice a distribuțiilor obținute permite o comparație a distribuțiilor de frecvențe obținute cu distribuția normală. Aceasta contează foarte mult pentru că, în măsura în care un poligon de frecvențe sau o histogramă diferă ca formă de curba normală, proporțiile ariilor de sub curbă nu mai sunt aceleași. Mai mult, modul particular în care distribuțiile diferă de curba normală pot avea implicații semnificative cu privire la date. Există multe moduri în care distribuțiile pot diferi de modelul curbei normale. Maniera și măsura în care se abat de la curba normală are implicații în ceea ce privește gradul de informații pe care distribuția îl transmite. Un caz extrem poate fi ilustrat de distribuția care ar rezulta dacă toate valorile dintr-un set de date ar apărea cu aceeași frecvență. O astfel de distribuție, care ar avea formă rectangulară, nu ar presupune nicio diferență a probabilității apariției unei valori date și, ca urmare, nu ar fi utilă pentru luarea deciziilor pe baza celor măsurate.

Un tip diferit, mai plauzibil, de abatere de la modelul curbei normale are loc atunci când distribuţiile au două sau mai multe moduri. Dacă o distribuţie este bimodală sau multimodală, trebuie să se ia în considerare posibilitatea problemelor de eşantionare sau caracteristicile speciale ale eşantionului. De exemplu, o distribuţie a calificativelor şcolare, în care frecvenţa cea mai mare apare pentru calificativele A şi D, cu puţine calificative B sau C, ar putea însemna că elevii din acea clasă sunt atipici într-un anume fel sau că ei aparţin unor grupuri care diferă semnificativ ca pregătire, motivaţie sau nivel al aptitudinii. Bineînţeles, informaţiile de această natură ar putea avea aproape invariabil implicaţii importante; în cazul acestui exemplu, ar putea determina

Info 2.5

De ce este atât de importantă curba normală în testarea psihologică?

În testare, modelul curbei normale este utilizat în moduri ce corespund distincţiei dintre statisticile descriptive şi inferenţiale:

1. Modelul curbei normale este utilizat *descriptiv* pentru a localiza poziţia scorurilor care provin din distribuţii normale. Într-un proces cunoscut ca *normalizare*, descris în Capitolul 3, curba normală este utilizată de asemenea pentru a face distribuţii care nu sunt normale — dar aproximează distribuţia normală — să fie similare modelului, în termenii poziţiilor relative ale scorurilor.
2. Modelul curbei normale este aplicat *inferenţial* în domeniul (a) *fidelităţii*, pentru a obţine intervalele de încredere necesare evaluării scorurilor obţinute şi a diferenţelor dintre scorurile obţinute (vezi Capitolul 4) şi (b) *validităţii*, pentru a obţine intervalele de încredere pentru predicţiile sau estimările bazate pe scorurile obţinute la test (vezi Capitolul 5).

un profesor să împartă clasa în secţiuni şi să utilizeze abordări pedagogice diferite pentru fiecare dintre acestea.

Două alte moduri în care distribuţiile se pot abate de la modelul curbei normale pot avea implicaţii semnificative relevante pentru datele obţinute la test. Aceste deviaţii ţin de proprietăţile de aplatizare şi asimetrie ale distribuţiilor de frecvenţe.

Indicele de aplatizare (Kurtosis)

Acest termen mai degrabă ciudat, care derivă din cuvântul grecesc pentru *convexitate*, se referă pur şi simplu la forma plată sau ascuţită a unei distribuţii. Distribuţiile de tip *platikurtic* au cel mai mare grad de dispersie, manifestată prin cozi mai extinse, iar distribuţiile de tip *leptokurtic* au cel mai mic grad de dispersie. Distribuţia normală este de tip mezokurtic, însemnând că are o dispersie de grad intermediar.

Aplicarea indicelui de aplatizare : Ipoteza variabilităţii mai mari în cazul bărbaţilor. În domeniul psihologiei diferenţiale, o ipoteză susţinută mult timp afirma că intervalul de inteligenţă este mai mare în cazul bărbaţilor decât în cazul femeilor. Ipoteza a apărut din observaţii cu privire la supra-reprezentativitatea bărbaţilor printre oamenii cu realizări extraordinare, ca şi în instituţiile pentru persoanele cu retard mintal. Deşi această ipoteză a fost dezbătută mult şi unii o susţin (vezi Halpern, 1997; Hedges & Nowell, 1995), din diferite motive — inclusiv natura testelor de inteligenţă — chestiunea nu este definitivată. Dacă ipoteza unei mai mari variabilităţi a bărbaţilor s-ar dovedi a fi exactă, ar însemna că se găsesc mai mulţi bărbaţi decât femei în zonele extreme ale distribuţiei scorurilor la un test de inteligenţă. Într-un astfel de caz, distribuţia scorurilor la un test de inteligenţă pentru femei şi pentru bărbaţi ar diferi în privinţa indicelui de aplatizare. Reprezentările

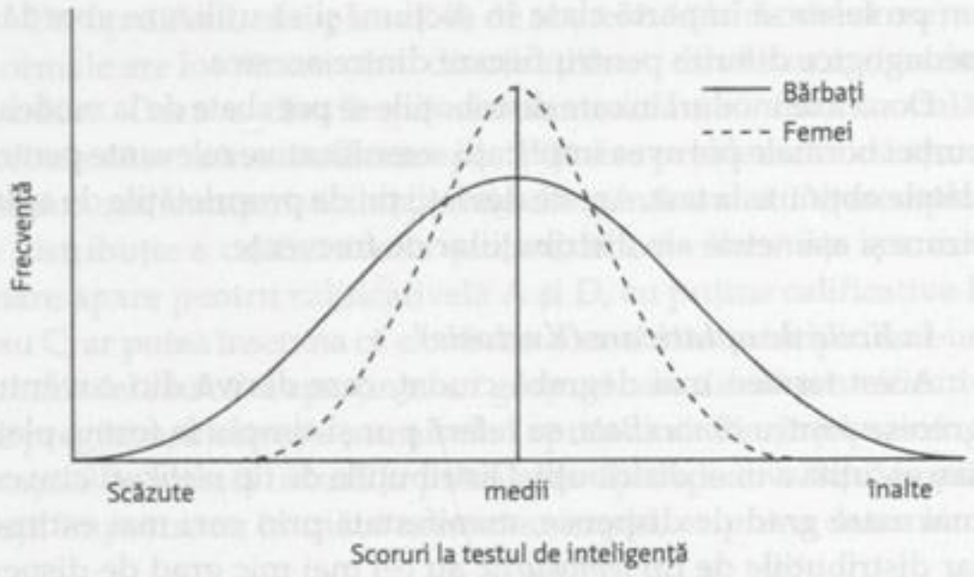

Figura 2.3 Distribuții ipotetice ale scorurilor la un test de inteligență, prezentând o variabilitate mai mare în cazul bărbaților decât în cazul femeilor (curbă de tip platikurtic vs. leptokurtic)

lor grafice, dacă ar fi suprapuse, ar putea arăta ca distribuțiile ipotetice ale scorurilor la un test de inteligență din Figura 2.3, care prezintă o distribuție de tip leptokurtic pentru femei și o distribuție de tip platikurtic pentru bărbați, fără nicio diferență între scorurile medii ale bărbaților și femeilor.

Indicele de asimetrie (Skewness)

Indicele de asimetrie (*Sk*) al unei distribuții se referă la o lipsă de simetrie. Așa cum am văzut, distribuția normală este perfect simetrică, cu $Sk = 0$; cea mai mare parte a valorilor se află la mijloc, iar cele două jumătăți sunt egale. Dacă cea mai mare parte a valorilor se află la capătul maxim al scalei și coada cea mai lungă

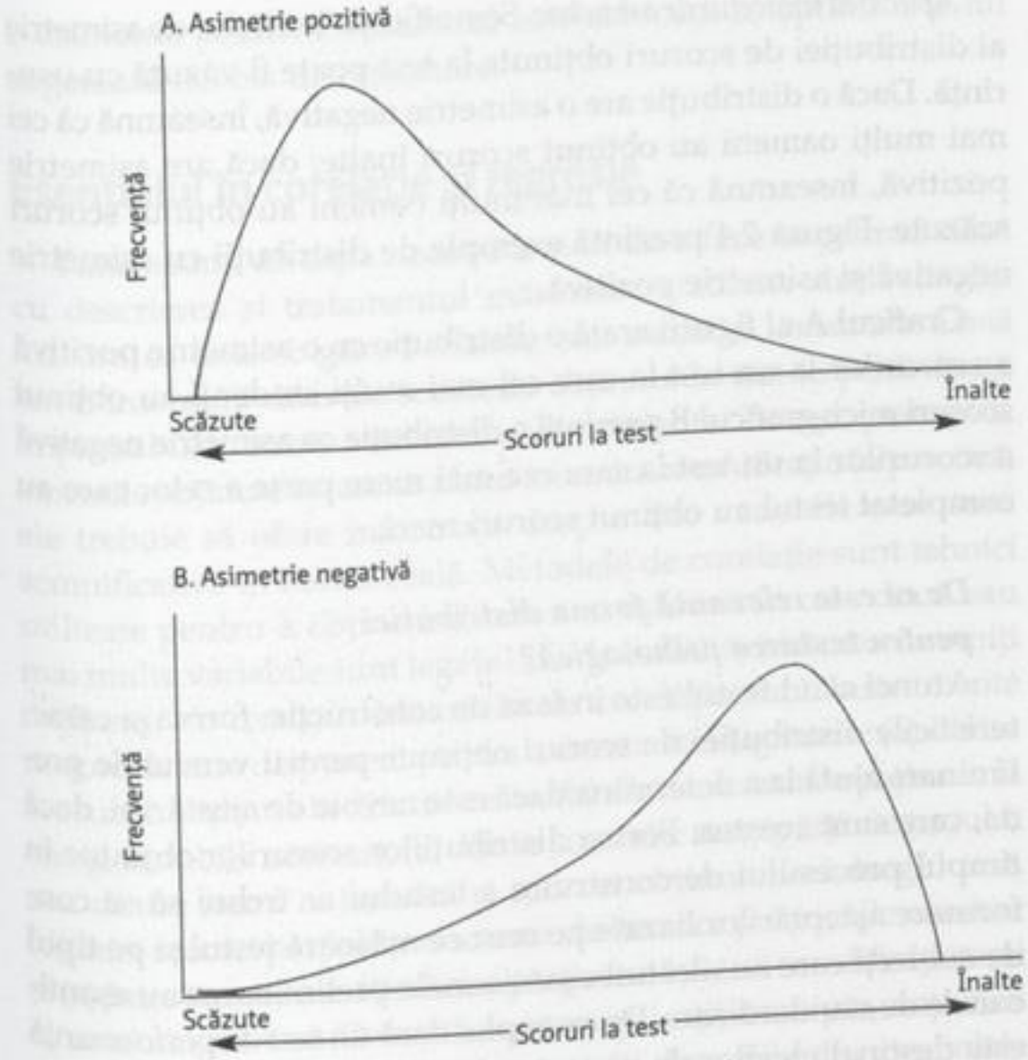

Figura 2.4. Distribuții asimetrice

se extinde către marginea inferioară a scalei, distribuția are o asimetrie *negativă* ($Sk < 0$); pe de altă parte, dacă cele mai multe dintre valori se află la marginea inferioară a scalei și coada mai lungă se extinde către marginea superioară a scalei, distribuția are asimetrie pozitivă ($Sk > 0$).

Aplicarea indicelui de asimetrie. Semnificația indicelui de asimetrie al distribuției de scoruri obținute la test poate fi văzută cu ușurință. Dacă o distribuție are o asimetrie negativă, înseamnă că cei mai mulți oameni au obținut scoruri înalte; dacă are asimetrie pozitivă, înseamnă că cei mai mulți oameni au obținut scoruri scăzute. Figura 2.4 prezintă exemple de distribuții cu asimetrie negativă și asimetrie pozitivă.

Graficul A al figurii arată o distribuție cu o asimetrie pozitivă a scorurilor la un test la care cei mai mulți studenți au obținut scoruri mici; graficul B prezintă o distribuție cu asimetrie negativă a scorurilor la un test la care cea mai mare parte a celor care au completat testul au obținut scoruri mari.

De ce este relevantă forma distribuției pentru testarea psihologică?

Atunci când testul este în faza de construcție, forma și caracteristicile distribuției de scoruri obținute pentru versiunile preliminare ajută la a determina dacă este nevoie de ajustări și, dacă da, care sunt acestea. Forma distribuțiilor scorurilor obținute în timpul procesului de construire a testului ar trebui să se conformeze așteptărilor bazate pe ceea ce măsoară testul și pe tipul de subiecți care au alcătuit eșantioanele preliminare sau eșantioanele de standardizare. De exemplu, dacă un test de performanță este destinat elevilor de liceu, iar distribuția scorurilor obținute într-un eșantion reprezentativ de elevi de liceu are o asimetrie negativă, înseamnă că testul ar putea fi prea ușor și că cel care construiește testul ar trebui să adauge itemi mai dificili pentru a aduce cea mai mare parte a scorurilor către centrul distribuției.

Invers, dacă același test este aplicat unui eșantion reprezentativ de elevi de școală primară și distribuția scorurilor lor la test are

o asimetrie pozitivă, rezultatul este conform așteptărilor și nu sugerează nevoia de reajustare.

Esențialul în corelație și regresie

Până acum, discuția noastră a avut de-a face în primul rând cu descrierea și tratamentul mărimilor statistice derivate din măsurarea unei singure variabile, sau statistici *univariate*. Dacă am fi interesați doar de scorurile obținute la test (o posibilitate improbabilă), aceste statistici ar fi suficiente. Totuși, pentru ca scorurile obținute la test să aibă o anumită semnificație în practică, ele trebuie să ofere informații despre alte variabile care sunt semnificative în lumea reală. Metodele de corelație sunt tehnici utilizate pentru a obține indicatori ai măsurii în care două sau mai multe variabile sunt legate una de cealaltă, indicatori numiți *coeficienți de corelație*. Metodele de corelație sunt instrumente majore de care dispunem pentru a demonstra legăturile (a) dintre scorurile obținute la diferite teste, (b) dintre scorurile obținute la test și variabile non-test, (c) dintre scorurile unor părți ale testului sau dintre itemi ai testului și scoruri obținute la întregul test, (d) între scoruri parțiale sau scoruri ale itemilor și variabile non-test și (e) între scoruri ale diferitelor părți ale unui test sau diferiți itemi ai unui singur test. Datorită acestor aplicații multiple, noțiunea de corelație joacă un rol important în discuțiile din capitolele următoare despre fidelitate, validitate și construirea testului.

Corelația ne introduce în domeniul statisticilor *bivariate* sau *multivariate*. În loc să avem o singură distribuție a frecvențelor unei singure variabile măsurate, avem nevoie de cel puțin două seturi de măsurători sau observații efectuate asupra aceluiași grup de oameni (ex.: scorurile obținute la două teste diferite) sau perechi

de observații pentru două seturi de indivizi (ex.: scorurile unor perechi de gemeni la același test). Atunci când calculăm un coeficient de corelație pentru a descrie relația dintre două variabile, datele sunt organizate sub formă de distribuții bivariate, cum sunt cele prezentate în Tabelul 2.6 pentru două seturi de date fictive.

Pentru a calcula un coeficient de corelație avem nevoie doar de datele (i.e. observații) obținute pentru două variabile. Variabilele ar putea fi venitul anual și anii de școală pentru un set de oameni, cantitatea de ploaie și mărimea plantelor pentru o perioadă de câțiva ani, lungimea medie a fustelor femeilor și performanța vânzărilor pe parcursul unei perioade de timp, poziția Soarelui pe cer și gradul de lumină de zi într-o locație dată, scorurile la un test și indicatorul performanței în muncă pentru un grup de angajați sau oricare altele. Două variabile (de obicei etichetate X și Y) pot fi corelate utilizând una dintre cele câteva metode de corelație care diferă în funcție de tipul de date și tipul de relații cărora le sunt destinate.

Corelația liniară

Relația dintre două variabile este numită *liniară* atunci când direcția și rata schimbării unei variabile sunt constante în raport cu schimbările celeilalte variabile. Atunci când sunt reprezentate grafic, punctele datelor pentru acest tip de relație formează un contur elipsoidal uniform sau aproape uniform. Dacă există o corelație între două variabile și relația dintre ele este liniară, există doar două posibilități: (a) o corelație pozitivă sau (b) o corelație negativă. Dacă nu există nicio corelație, punctele datelor nu se aliniază într-o formă sau tendință definită și putem presupune că cele două seturi de date nu au o sursă comună a dispersiei.

Tabelul 2.6 Două seturi de date bivariate

Individ	Scor la testul X	Scor la testul Y
	A. Date pentru o corelație pozitivă perfectă	
1	3	5
2	4	6
3	6	8
4	7	9
5	8	10
6	9	11
7	10	12
8	11	13
9	13	15
10	14	16
	B. Date pentru o corelație negativă perfectă	
1	140	5
2	130	6
3	110	8
4	100	9
5	90	10
6	80	11
7	70	12
8	60	13
9	40	15
10	30	16

Dacă există fie o corelație pozitivă, fie o corelație negativă, de orice mărime, putem evalua posibilitatea ca acea corelație să fi apărut datorită întâmplării utilizând mărimea eșantionului pe care a fost calculată corelația și tabelele statistice care prezintă probabilitatea

ca un coeficient de o mărime dată să apară din întâmplare. Dacă probabilitatea ca acel coeficient obținut să rezulte din întâmplare este foarte mică, putem avea încredere că există o corelație mai mare decât zero între X și Y. În astfel de cazuri presupunem că cele două variabile au un anumit grad de dispersie comună. Cu cât coeficientul de corelație este mai mare și mai semnificativ statistic, cu atât este mai mare gradul de dispersie despre care presupunem că este comun variabilelor X și Y. Proporția de dispersie comună a celor două variabile este deseori estimată prin pătratul coeficientului de corelație (r_{xy}) prin care se obține *coeficientul de determinare* sau r^2_{xy}. Deși coeficienții de determinare ne spun cât de mult din dispersia lui Y poate fi explicată prin dispersia lui X sau vice versa, ei nu indică în mod necesar faptul că există o relație cauzală între X și Y.

Diagrame de împrăștiere

Reprezentarea grafică a datelor bivariate sub forma diagramelor de împrăștiere sau a diagramelor prin nori de puncte este esențială pentru a vizualiza tipul de relație cu care aveam de-a face. Diagramele prin nori de puncte din figura 2.5 prezintă tiparul punctelor care rezultă din reprezentarea distribuțiilor bivariate din Tabelul 2.6. Aceste figuri ne fac să vedem în mod concret forța și direcția relației dintre cele două variabile din fiecare set de date. Putem vedea că în ambele părți ale figurii, tiparul urmează o linie diagonală, indicând faptul că ambele relații reprezentate sunt liniare și puternice; de fapt, corelațiile sunt perfecte, ceea ce rareori este întâlnit în cazul datelor reale. O corelație puternică înseamnă că, pe măsură ce valorile unei variabile cresc sau descresc, există un grad corespunzător de schimbare a valorilor celeilalte

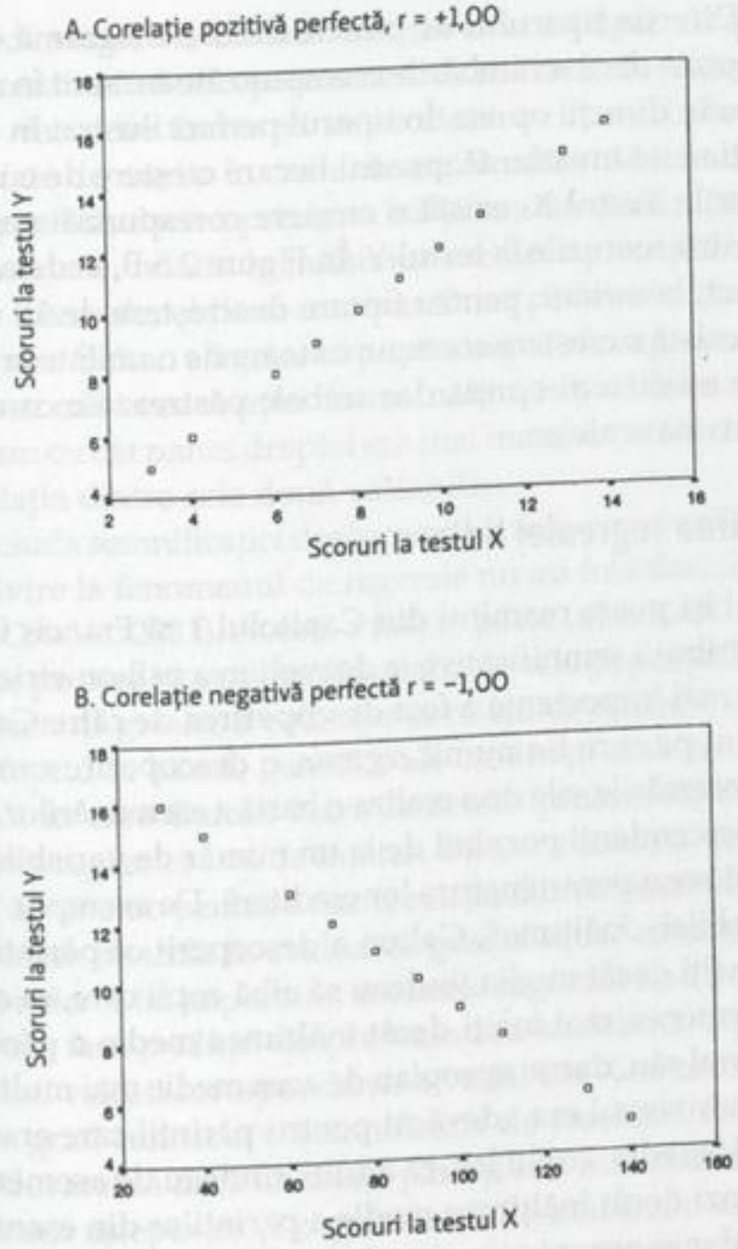

Figura 2.5 Diagrame prin nori de puncte ale datelor bivariate din tabelul 2.6

variabile. Direcția tiparului de puncte dintr-o diagramă de dispersie ne spune dacă schimbările corespunzătoare sunt în aceeași direcție sau în direcții opuse. În tiparul perfect ilustrat în Figura 2.5 A, relația este invariantă: pentru fiecare creștere de o unitate a scorurilor la Testul X, există o creștere corespunzătoare de o unitate pentru scorurile la testul Y. În Figura 2.5 B, vedem un alt tipar perfect, invariant: pentru fiecare descreștere de 10 unități la testul X există o creștere corespunzătoare de o unitate la Testul Y. Relațiile au direcții opuse, dar ambele păstrează o corespondență relativă a scalelor.

Descoperirea regresiei liniare

Cititorul își poate reaminti din Capitolul 1 că Francis Galton a adus contribuții semnificative la dezvoltarea psihometriei. Una dintre cele mai importante a fost descoperirea de către Galton a fenomenului pe care l-a numit *regresie*, o descoperire care a rezultat din încercările sale de a realiza o hartă a asemănărilor dintre părinți și descendenți pornind de la un număr de variabile și de a produce dovezi pentru natura lor ereditară. De exemplu, în termenii variabilei „înălțime", Galton a descoperit că părinții care erau mai înalți decât media tindeau să aibă copii care, ca adulți, erau de asemenea mai înalți decât înălțimea medie a părinților din eșantionul său, dar se apropiau de acea medie mai mult decât părinții lor. Inversul era adevărat pentru părinții care erau mai scunzi decât media; copiii lor, ca adulți, tindeau de asemenea să fie mai scunzi decât înălțimea medie a părinților din eșantionul lui Galton, dar se apropiau mai mult de acea medie decât părinții lor. Când a reprezentat grafic aceste date bivariate ale seturilor corespondente ale înălțimilor părinților și copiilor, precum și alte

seturi de variabile, Galton a sesizat că acest tipar al *regresiei către medie* continuă să se repete: scorurile extreme ale părinților la una dintre variabile tindea să se asocieze cu scoruri care erau mai apropiate de medie în cazul descendenților. Mai mult, Galton a descoperit că, dacă reprezenta înălțimea descendenților din diferite intervale, relativ la înălțimea medie a părinților din cadrul acelor intervale, obținea un tipar liniar pe care l-a numit *dreaptă de regresie*. Galton a înțeles că panta dreptei de regresie reprezenta forța sau mărimea relației dintre înălțimile părinților și înălțimile copiilor: cu cât panta dreptei era mai mare, cu atât mai puternică era relația dintre cele două variabile.

În ciuda semnificației descoperirilor sale, concluziile lui Galton cu privire la fenomenul de regresie nu au fost destul de precise (vezi Cowles, 2001). Acesta a fost în parte un rezultat al limitelor datelor pe care le-a utilizat în analizele sale și, parțial, s-a datorat interpretării greșite a cauzelor corelațiilor dintre variabile. Dat fiind faptul că bazele genetice ale eredității erau neclare în momentul în care Galton lucra la aceste probleme, interpretarea greșită a regresiei este de înțeles. Cu toate acestea, procedura pe care a inventat-o pentru a desrie relația dintre variabile s-a dovedit a fi extrem de utilă în evaluarea gradului de varianță comună a variabilelor. Mai important, analizele regresiei ne-au oferit o bază pentru a face *predicții* despre valoarea variabilei Y pe baza cunoașterii valorii corespondente a variabilei X cu care variabila Y are un grad cunoscut și semnificativ de corelație. Până la urmă, este adevărat că dacă un set de părinți este mai înalt decât media, ne putem aștepta ca și copiii lor să fie mai înalți decât media. Galton însuși a inventat un mod de a cuantifica relația dintre variabile, transformând valorile fiecăreia pe o scală comună și calculând un indice sau coeficient numeric care însumează forța

relației lor. Totuși, Karl Pearson, matematician și discipol al lui Galton, este cel care a rafinat metoda și a construit formula cel mai larg utilizată de calcul al coeficienților de corelație.

Coeficienți de corelație

Așa cum am văzut, *corelația* se referă pur și simplu la măsura în care variabilele sunt legate unele de altele. Gradul și direcția corelației dintre variabile sunt măsurate prin intermediul diferitelor tipuri de *coeficienți de corelație*, care sunt numere ce pot fluctua între -1,00 și +1,00. Info 2.6 enumeră câteva dintre apectele fundamentale, dar deseori înțelese greșit, ale coeficienților de corelație în general.

Spre deosebire de așa-numitele științe „tari", în care experimentul este modalitatea tipică de a progresa, în științele comportamentului capacitatea de a manipula variabilele este deseori limitată. Astfel, cercetarea în psihologie se bazează într-o mare măsură pe metode corelaționale. Din fericire, aria proiectelor de cercetare și a metodelor de analiză ce pot fi aplicate datelor a crescut imens odată cu creșterea puterii și disponibilității computerelor moderne. Unele dintre tehnicile utilizate în mod curent pentru analiza simultană a datelor provenite de la mai multe variabile, cum ar fi regresia multiplă și analiza „path", sunt atât de sofisticate, încât le permit psihologilor și altor oameni de știință din domeniul social să facă anumite inferențe cu privire la relațiile cauzale cu un nivel înalt de încredere.

Tehnica statistică utilizată pentru a calcula un coeficient de corelație depinde de natura variabilelor care trebuie corelate, tipurile de scale utilizate în măsurarea lor și tiparul relației dintre acestea. Din nou, o trecere în revistă completă a metodelor depășește

Info 2.6

Trei aspecte esențiale ale corelației în general

1. *Gradul relației dintre două variabile este indicat de numărul din coeficient, în timp ce direcția relației este indicată de semn.* De exemplu, un coeficient de corelație de -0,80 indică exact același grad al relației ca și un coeficient +0,80. Indiferent dacă este pozitivă sau negativă, o corelație este redusă în măsura în care coeficientul său se apropie de zero. Deși aceste fapte pot părea evidente, natura restrictivă a semnului negativ face ca mulți oameni să le uite.
2. *Corelația, chiar dacă este mare, nu presupune cauzalitate.* Dacă două variabile, X și Y, sunt corelate, poate fi pentru că X este cauza lui Y, pentru că Y este cauza lui X sau pentru că o a treia variabilă Z este cauză atât pentru X, cât și pentru Y. Acest truism este deseori frecvent ignorat; coeficienți de corelație situați între moderat și înalt sunt deseori prezentați ca și cum ar fi dovada unei relații cauzale între variabilele corelate.
3. *Corelațiile înalte ne permit să facem predicții.* Deși corelația nu presupune cauzalitate, presupune un anumit grad de variație comună. Cunoașterea măsurii în care lucrurile variază în relație unul cu celălalt este extrem de utilă. Prin analiza regresiei putem utiliza datele corelaționale ale uneia sau mai multor variabile pentru a obține ecuații care să ne permită să prezicem valori așteptate ale unei variabile dependente (Y) într-un anumit interval de eroare, pe baza valorilor cunoscute ale uneia sau mai multor variabile (X_1, X_2,... X_k) cu care variabila dependentă corelează.

scopul acestei cărți. Cu toate acestea, indicatorii cei mai frecvent utilizați ai gradului de corelație dintre două variabile vor fi luați în discuție.

Coeficientul de corelație produs-moment Pearson

Formula fundamentală introdusă de Karl Pearson pentru calcularea coeficientului de corelație al datelor bivariate dintr-un eșantion este cunoscută în mod formal drept *coeficient de corelație produs-moment Pearson*. Formula de definire a acestui coeficient, numit în mod obișnuit Pearson *r*, este

$$r_{xy} = \frac{\Sigma\, xy}{N s_x s_y} \qquad (2.1)$$

unde

r_{xy} = corelația dintre X și Y;
x = abaterea unui scor X de la media scorurilor X
y = abaterea unui scor Y corespunzător de la media scorurilor Y;
Σ xy = suma tuturor produselor încrucișate ale abaterilor (i.e., suma produselor numărului de abateri ale fiecărui x cu abaterea y corespondentă);
N = numărul de perechi dintr-un set de date bivariate;
s_x = abaterea standard a scorurilor X
s_y = abaterea standard a scorurilor Y

Deși formula de calcul al scorurilor brute pentru coeficientul de corelație Pearson *r* este mai complicată decât formula de definiție, ușurința accesului la programele pe calculator pentru calcularea coeficienților de corelație face ca formulele de calcul să nu fie necesare. Pe de altă parte, Formula (2.1) și chiar și Formula (2.2), mai scurtă, sunt de un ajutor considerabil în înțelegerea semnificației numerice a coeficientului de corelație.

Info 2.7 enumeră notațiile de bază pentru corelație, alături de două versiuni ale formulei pentru coeficientul Pearson *r*.

Coeficientul Pearson *r* este de fapt media produselor încrucișate ale scorurilor standard ale celor două variabile corelate. Formula care concretizează această definiție este

$$r_{xy} = \frac{\Sigma\, z_x z_y}{N} \qquad (2.2)$$

unde

r_{xy} = corelația dintre X și Y
z_x = scorurile standard ale variabilei x, obținute prin împărțirea fiecărui scor de abatere al abaterea standard a lui X; și
z_y = scorurile standard ale variabilei z, obținute prin împărțirea fiecărui scor de abatere al abaterea standard a lui Y

Suma produselor încrucișate ale scorurilor z ale variabilelor X și Y și împărțirea la numărul de perechi dintr-un set de date duce la o medie care reflectă mărimea relației dintre X și Y, și anume coeficientul Pearson *r*.

Formula (2.2) prezintă interes în contextul testării psihologice nu doar datorită conciziei sale și bazei conceptuale, ci și pentru că servește la introducerea noțiunii de *scoruri standard*, sau *scoruri z*, cu care ne vom întâlni din nou în capitolul următor. S-ar putea ca cititorul să fi observat că valorile de-a lungul abscisei curbei normale din Tabelul ariilor situate sub curba normală, prezentat în Anexa C, sunt date în termeni de scoruri z. Motivul este că un scor z reprezintă distanța dintre fiecare valoare dintr-o distribuție și media acelei distribuții, exprimate în termeni de

Info 2.7

Notații de bază ale corelației

Variabila X	Variabila Y
X= un scor al variabilei X	Y= un scor al variabilei Y
$x = X - \bar{X}$ scorul abaterii variabilei X	$y = Y - \bar{Y}$ scorul abaterii variabilei X
S_x = Abaterea standard a lui X	S_y = Abaterea standard a lui Y
z_x= Scor standard al variabilei X	z_y= Scor standard al variabilei Y
$z_X = \frac{X - \bar{X}}{s_x}$	$z_Y = \frac{Y - \bar{Y}}{s_y}$

Formulele pentru Pearson *r*:

$$r_{xy} = \frac{\Sigma xy}{Ns_x s_y}$$ Formula (2.1), formula de definiție

$$r_{xy} = \frac{\Sigma z_x z_y}{N}$$ Formula (2.2), formula cu scoruri standard

Unde *N* = numărul de observații pereche ale lui X și Y utilizate pentru a calcula *r*.

Coeficientul de determinare = r_{xy}^2

unități ale abaterii standard pentru acea distribuție. Formula care utilizează scoruri standard (2.2) pentru coeficientul Pearson *r* oferă pur și simplu o cale mai concentrată de exprimare a relației dintre două variabile.

Condiții necesare pentru utilizarea coeficientului Pearson *r*

Deși este cel mai larg utilizat coeficient de corelație, Pearson *r* este potrivit numai pentru date care îndeplinesc câteva condiții. De când Pearson a introdus această formulă, s-au construit diferite metode de a obține coeficienți de corelație pentru diferite tipuri de date bivariate. Obținerea coeficientului de corelație produs-moment se bazează pe următoarele premise:

1. Perechile de observații sunt independente una de alta.
2. Variabilele de corelat sunt continue și măsurate pe scale de interval sau rată.
3. Relația dintre variabile este liniară; adică, ea aproximează o dreaptă, așa cum am descris mai sus.

Dacă sunt îndeplinite prima și a doua dintre aceste premise sau condiții se stabilește cu ușurință prin cunoașterea manierei în care au fost culese datele și a tipului de date manevrate. Dacă perechile de scoruri sau de observații de corelat sunt obținute în mod independent una de alta, prima premisă este satisfăcută. Dacă datele ambelor variabile reprezintă cantități continue, cea de-a doua premiză este îndeplinită.

Satisfacerea celei de-a treia premise a coeficientului Pearson *r*, cea mai critică, necesită examinarea diagramei prin nori de puncte a datelor bivariate pentru a vedea dacă distribuția cazurilor are forma elipsoidală care este un indicator al relației liniare prezentate în Figura 2.6 A. Atunci când această premisă este încălcată, coeficientul Pearson *r* nu este un indicator precis al corelației.

Deviații de la liniaritate

În scopul determinării aplicabilității coeficientului Pearson *r* unui set de date bivariate, există două moduri în care un nor de puncte poate devia de la forma elipsoidală care indică o relație liniară pozitivă. Primul și cel mai evident mod este existența unei arcuiri semnificative a formei elipsoidale, ca în Figura 2.6, B și C. Astfel de deviații indică faptul că nu există nicio relație liniară și, prin urmare, că relația dintre X și Y nu este aceeași pe tot intervalul valorilor lor. A doua modalitate în care norul de puncte poate devia de la elipsa ce indică o relație liniară este o condiție numită *heteroscedasticitate*. Aceasta înseamnă, pur și simplu, că dispersia sau variabilitatea din diagrama prin nor de puncte nu este uniformă pe intervalul de valori ale celor două variabile. Pentru a utiliza coeficientul de corelație Pearson *r*, norul de puncte trebuie să prezinte un grad relativ uniform al dispersiei, sau *homoscedasticitate*, în acel interval. Diagrama prin nor de puncte din Figura 2.6 A este homoscedastică, în timp ce Figura 2.6 D și E sunt heteroscedastice. Unul dintre modurile de a evita aplicarea neadecvată a coeficientului Pearson *r* este realizarea unei diagrame prin nor de puncte a datelor bivariate și examinarea formei sale, urmărind posibile deviații de la liniaritate. Dacă formula coeficientului Pearson *r* este aplicată datelor care deviază de la o relație strict liniară, fie în termenii arcuirii formei norului de puncte sau datorită heteroscedasticității, coeficientul de corelație rezultat va fi un indicator incorect al relației dintre X și Y.

Restricții de interval și corelație

O trăsătură importantă și deseori trecută cu vederea a coeficientului Pearson *r* se referă la modul în care este influențat de variabilitatea variabilelor corelate. Formulat în mod simplu,

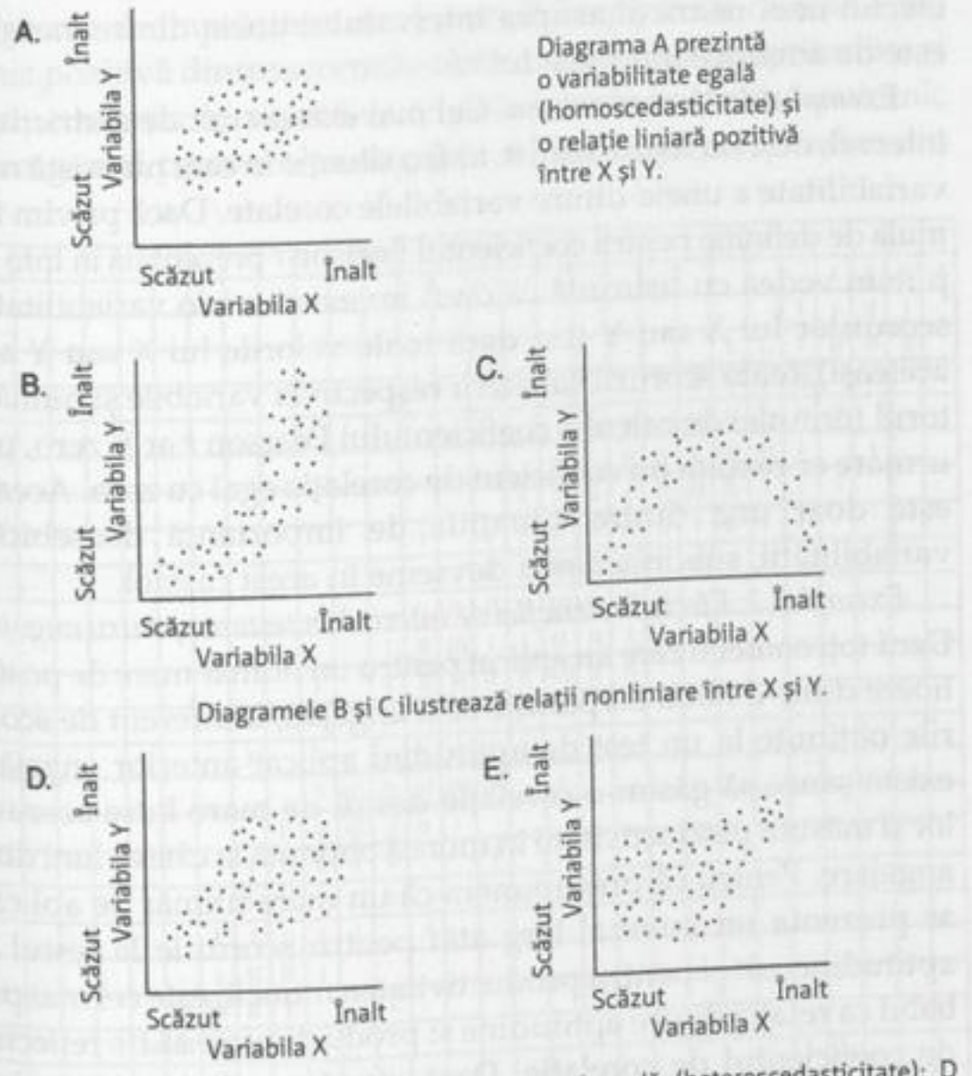

Figura 2.6 Diagrame prin nori de puncte care ilustrează diferite caracteristici ale datelor bivariate.

Notă: Fiecare punct marchează localizarea unei perechi de observații, sau scoruri la X și Y.

efectul unei restricții asupra intervalului uneia dintre variabile este de a reduce mărimea lui *r*.

Exemplul 1: Un caz extrem. Cel mai extrem caz de restricție de interval, deși nu foarte realist, ar fi o situație în care nu există nicio variabilitate a uneia dintre variabilele corelate. Dacă privim formula de definiție pentru coeficientul Pearson *r* prezentată în Info 2.7, putem vedea cu ușurință că dacă nu există nicio variabilitate a scorurilor lui X sau Y (i.e. dacă toate valorile lui X sau Y ar fi aceleași), toate scorurile abaterii respectivei variabile și numărătorul formulei de calcul a coeficientului Pearson *r* ar fi zero, prin urmare ar rezulta un coeficient de corelație egal cu zero. Aceasta este doar una dintre situațiile de importanță deosebită a variabilității, subliniată mai devreme în acest capitol.

Exemplul 2: Efectul restricției de interval în testarea pentru angajare. Dacă toți oamenii care au aplicat pentru un număr mare de posturi libere dintr-o nouă corporație ar fi angajați, indiferent de scorurile obținute la un test de aptitudini aplicat anterior angajării, există șanse să găsim o corelație destul de mare între scorurile lor și măsura productivității în muncă obținută la câteva luni după angajare. Pentru că presupunem că un mare număr de aplicații ar prezenta un interval larg atât pentru scorurile la testul de aptitudine, cât și pentru productivitatea muncii, este cel mai probabil ca relația dintre aptitudine și productivitate să fie reflectată de coeficientul de corelație. Dacă, după o vreme, procesul de selecție a personalului ar deveni mai restrictiv — astfel încât doar acei candidați care obțin scoruri înalte vor fi angajați — efectul net al acestei schimbări va trebui să restrângă intervalul aptitudinii pentru noii angajați. Astfel, dacă s-ar calcula un nou coeficient doar cu datele noilor angajați, gradul de corelație dintre scorurile obținute la testul de aptitudine și productivitate va fi mai redus.

Diagrama de împrăștiere din Figura 2.7 reprezintă corelația puternic pozitivă dintre scorurile obținute la testul de aptitudine și productivitatea în muncă ce ar putea exista în grupul puternic heterogen al persoanelor angajate inițial. Segmentul redus din

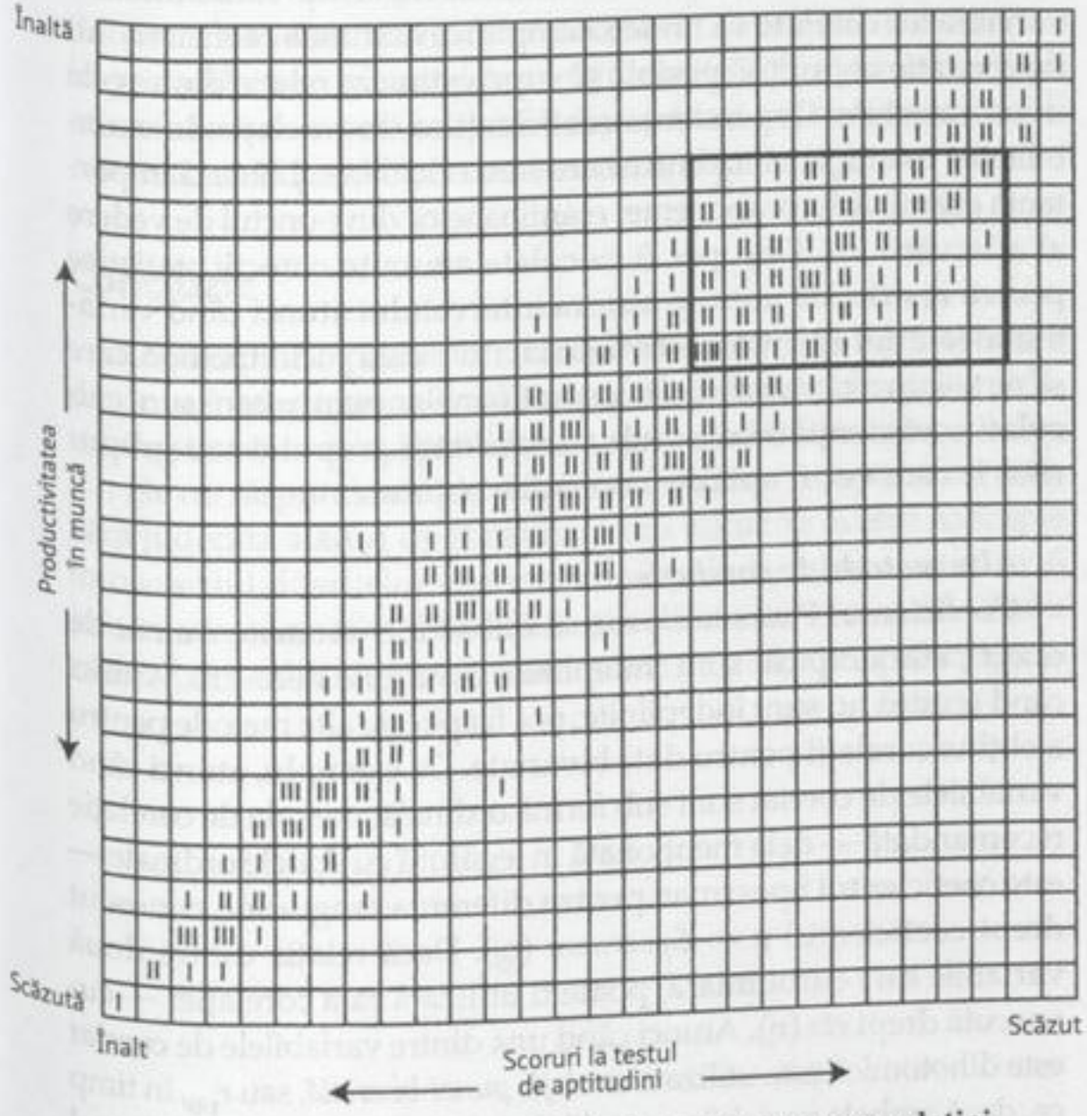

Figura 2.7 Efectul restricțiilor de interval asupra corelației

partea extremă dreaptă a diagramei reprezintă o corelație redusă, aproape inexistentă, care ar putea exista în grupul mult mai limitat al candidaților superiori, angajați mai târziu.

Așa cum o restricționare a intervalului variabilelor corelate va scădea corelația dintre ele, un interval mare de variabilitate a variabilelor corelate va tinde să amplifice mărimea coeficientului de corelație obținut și, posibil, să supraestimeze relația dintre cele două variabile. Dependența coeficienților de corelație de variabilitatea eșantioanelor pentru care sunt calculați subliniază importanța examinării componenței eșantioanelor din punctul de vedere al adecvării lor. Deși pot fi calculate anumite corecții statistice pentru restricțiile impuse asupra intervalului atunci când variabilitatea unui eșantion este redusă, nu există nicio metodă care să ne asigure că variabilitatea eșantioanelor pentru care sunt calculați coeficienții corespunde variabilității grupului sau grupurilor la care vor fi aplicate corelațiile obținute.

Alte metode de corelație

Coeficientul Pearson *r* poate fi folosit într-un mare număr de cazuri, atât timp cât sunt îndeplinite condițiile necesare. Atunci când acestea nu sunt îndeplinite, pot fi aplicate alte metode pentru a obține corelații pentru date bivariate. De exemplu, atunci când variabilele de corelat sunt sub formă ordinală, metoda de corelație recomandată — deja menționată în legătură cu scalele ordinale — este coeficientul Spearman pentru diferența rangurilor, cunoscut drept coeficientul ρ — *Spearman* (r_S). Dacă relația dintre două variabile este curbiliniară, poate fi utilizată rata corelației — cunoscută drept *eta* (η). Atunci când una dintre variabilele de corelat este dihotomică este utilizată *corelația punct-biserială*, sau r_{pb}, în timp ce, dacă ambele variabile sunt dihotomice, este folosit coeficientul

phi (ϕ). Variabilele dihotomice intră în joc deseori în analiza de itemi a datelor unui test înregistrate în termeni de răspunsuri corect — greșit sau adevărat — fals.

Există multe alte tipuri de coeficienți de corelație care sunt adecvați unor tipuri specifice de date. Ei pot fi găsiți în manualele de statistică atunci când este necesar. O variantă deosebit de importantă este *coeficientul de corelație multiplă* (*R*), care este utilizat atunci când o singură variabilă dependentă (Y) corelează cu doi sau mai mulți predictori combinați ($X_1, X_2, \ldots, X_k$).

Concluzie

Acest capitol a prezentat conceptele statistice de bază necesare pentru a înțelege scorurile obținute la test și semnificația lor. Mărimile statistice există pentru a ne ajuta să dăm un sens datelor, dar ele nu răspund la întrebări. Pentru aceasta trebuie să ne utilizăm judecata alături de statistici. Vom reîntâlni aceste concepte în contextul diferitelor aspecte tehnice ale testelor — cum ar fi informația normativă, fidelitatea și validitatea — care ne permit să evaluăm calitatea lor ca instrumente de măsurare psihologică.

TESTEAZĂ-TE SINGUR

1. **Care dintre următoarele scale de măsură este singura care are un punct zero semnificativ?**
   (a) Nominală
   (b) Raport
   (c) Ordinală
   (d) Interval

2. **Tom și Jerry au obținut scoruri în percentilele 60 și 65 la un test de aptitudini verbale. Mary și Martha au obținut scoruri în percentilele 90 și 95 la același test. Putem concluziona că diferența dintre Tom și Jerry în termenii aptitudinilor lor verbale este aceeași cu diferența dintre Mary și Martha. Afirmația precedentă este**
   (a) adevărată.
   (b) falsă.

3. **La un test de aptitudini cognitive generale, un copil în vârstă de 5 ani obține un scor al vârstei mentale de 4 ani și un copil în vârstă de 10 ani obține un scor al vârstei mentale de 9 ani. Dacă li s-ar calcula IQ-ul conform formulei originale, rezultatul ar fi următorul:**
   (a) Ambii copii ar obține același raport IQ.
   (b) Copilul de 5 ani ar obține un raport IQ mai mare.
   (c) Copilul de 10 ani ar obține un raport IQ mai mare.

4. **În distribuția 2, 2, 2, 2, 3, 3, 3, 8, 11, media, mediana și modul sunt, respectiv,**
   (a) 4, 3 și 2.
   (b) 2, 4 și 3.
   (c) 3, 4 și 2.
   (d) 2, 3 și 4.

5. **Pentru testare și pentru multe alte scopuri, cel mai tipic indicator al variabilității scorurilor unei distribuții este**
   (a) suma pătratelor scorurilor abaterii.
   (b) rădăcina pătrată a dispersiei.
   (c) amplitudinea semi-intercuartilică.

6. **Care dintre următoarele afirmații despre modelul curbei normale nu este adevărată?**
   (a) Este simetrică bilateral.
   (b) Limitele sale se extind la infinit.
   (c) Media, mediana și modul coincid.
   (d) Este multimodală.

7. **Aria unei distribuții normale cuprinse între +1σ și –1σ reprezintă aproximativ _____ din aria de sub curbă.**
   (a) 50%
   (b) 68%
   (c) 95%
   (d) 99%

8. **Dacă forma distribuției scorurilor obținute la un test este semnificativ asimetrică, înseamnă că testul este probabil _____ pentru cei cărora le-a fost aplicat.**
   (a) prea ușor
   (b) prea greu
   (c) fie prea ușor, fie prea greu
   (d) adecvat

9. **Care dintre următorii coeficienți reprezintă cel mai puternic grad de corelație dintre două variabile?**
   (a) – 0,80
   (b) – 0,20
   (c) + 0,20
   (d) + 0,60

10. **Dacă intervalul de valori al oricăreia dintre două variabile corelate utilizând coeficientul de corelație moment-produs Pearson (Pearson *r*) este restricționat, mărimea coeficientului obținut va fi**
    (a) mai mică.
    (b) mai mare
    (c) neafectată.

Răspunsuri corecte: 1. b; 2. b; 3. c; 4. a; 5. b; 6. d; 7. b; 8. c; 9. a; 10. a

Aspecte esențiale ale interpretării scorurilor obținute la test

Indiferent câte statistici sunt utilizate în testarea psihologică, în analiza finală semnificația scorurilor la un test derivă din cadrul de referință pe care-l utilizăm pentru a le interpreta și din contextul în care sunt obținute scorurile. Pentru a fi siguri, scorurile la test au nevoie, de asemenea, să fie valide și itemii testului trebuie să fie construiți și evaluați cu atenție, astfel încât să contribuie la semnificația scorurilor; acestea sunt aspecte de care ne vom ocupa în Capitolele 4 și 6. În acest capitol ne îndreptăm atenția către cadrele de referință pentru interpretarea scorurilor testului, o temă strâns legată de validitatea inferențelor pe care le putem face pe baza testelor, discutată pe larg în Capitolul 5. În contextul în care are loc testarea, un aspect de o importanță deosebită legat de procesul de selecție a testului și de administrarea lui este discutat în capitolul final. Info 3.1 enumeră trei surse excelente de informație în care multe dintre subiectele discutate în acest capitol sunt parcurse în amănunțime.

Scorurile brute

Un *scor brut* este un număr (X) care însumează sau surprinde un anumit aspect al performanței unei persoane la un eșantion de comportament atent selectat și observat care alcătuiește testul psihologic. În sine, un scor brut nu transmite nicio semnificație. Scorurile mari pot reprezenta un rezultat favorabil la un test de aptitudine, dar nefavorabil la un test care evaluează un anumit aspect al psihopatologiei. De exemplu, în Inventarul Multifazic de Personalitate Minnesota (MMPI) scorurile ridicate indică de obicei un anumit tip de dezadaptare, în timp ce scorurile scăzute nu indică în mod necesar o bună adaptare. Chiar dacă știm tipul de test la care a fost obținut un scor, putem fi pe o cale greșită. Unele teste de aptitudini cognitive — în special numeroase instrumente neuropsihologice — sunt scorate în termenii numărului de erori sau de viteză a performanței, astfel încât cu cât este mai mare scorul, cu atât mai puțin favorabil este rezultatul. Mai mult, nu putem ști cât de înalt este „înalt" fără niciun cadru de referință. Un scor care pare a fi înalt — cum ar fi un IQ de 130, de exemplu — poate avea semnificații destul de diferite, în funcție de testul din care a fost obținut, ariile pe

Info 3.1

Pentru informații mai extinse despre aspectele tehnice ale unora dintre subiectele discutate în acest capitol, vezi una dintre următoarele surse:

- Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.
- Petersen, N. S., Kolen, M. J. & Hoover, H. D. (1989). Scaling, norming, and equating. În R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: American Council on Education/Macmillan.
- Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Mahwah, NJ: Erlbaum.

care testul le acoperă, cât de recente sunt normele sale și în funcție de aspectele specifice ale situației în care scorul a fost obținut și de caracteristicile persoanei căreia i s-a aplicat testul.

Cadrele de referință pentru interpretarea scorurilor obținute la test

La baza tuturor celorlalte aspecte privind interpretarea scorurilor, într-un mod sau altul, se află problema cadrelor de referință utilizate pentru a interpreta un scor dat. În funcție de scopurile lor, testele se bazează pe una dintre următoarele surse de informație pentru a obține cadre de referință pentru semnificația lor:

1. *Norme. Interpretarea testului în raport cu norma* utilizează standarde bazate pe performanța unor grupuri specifice de oameni care oferă informații pentru interpretarea scorurilor. Acest tip de interpretare a testului este util în primul rând atunci când trebuie să comparăm indivizii unul cu altul sau cu un grup de referință pentru a evalua diferențele dintre ei cu privire la o caracteristică măsurată de test. Termenul *normă* se referă la performanța la test sau la comportamentul tipic al unuia sau mai multor grupuri de referință. Normele sunt de obicei prezentate sub forma tabelelor cu statistici descriptive — cum ar fi medii, abateri standard și distribuții de frecvență — care însumează performanța grupului sau grupurilor în chestiune. Atunci când normele sunt extrase din performanța la test a grupurilor de oameni, aceste grupuri de referință sunt etichetate drept *eșantioane normative* sau *de standardizare*. Obținerea normelor este un aspect central al procesului de standardizare a unui test cu raportare la normă.
2. *Criteriul de performanță.* Atunci când relația dintre itemii sau sarcinile testului și standardul performanței este demonstrabilă și bine definită, scorurile obținute la test pot fi evaluate via *interpretare în raport cu un criteriu*. Acest tip de interpretare utilizează proceduri, cum ar fi eșantionarea domeniilor de conținut sau a comportamentelor legate de activitatea de muncă, proiectate pentru a stabili dacă și în ce măsură nivelele dorite ale criteriilor de măiestrie sau de performanță au fost îndeplinite.

Interpretarea unui test prin raportare la normă

De departe, normele constituie cel mai larg utilizat cadru de referință pentru interpretarea scorurilor obținute la test. Performanța unor grupuri bine stabilite de oameni este utilizată ca bază pentru interpretarea scorurilor atât în testarea aptitudinilor, cât și în testarea personalității. Atunci când normele reprezintă cadrul de referință, întrebarea la care acestea răspund în mod tipic este „Cum este performanța celui testat în comparație cu performanța celorlalți?" Scorul este utilizat pentru a plasa performanța subiectului testat în cadrul unei distribuții preexistente a scorurilor sau a datelor obținute din performanța unui grup de comparație potrivit.

Norme de dezvoltare

Scale la nivel ordinal bazate pe secvențe comportamentale

Dezvoltarea umană este caracterizată de procesul secvențial al unui număr de arii de comportament. Un exemplu clasic este

secvența parcursă de dezvoltarea motorie în timpul perioadei de sugar. În primul an de viață, cei mai mulți copii progresează de la postura fetală la naștere, la a sta în șezut și în picioare, până la a merge singuri. Ori de câte ori o secvență universală de dezvoltare presupune o progresie ordonată de la un stadiu comportamental la alt stadiu — mai avansat — secvența poate fi convertită într-o scală *ordinală* și utilizată normativ. În astfel de cazuri, cadrul de referință pentru interpretarea scorurilor testului este derivat din observarea și notarea anumitor regularități în ordinea și temporizarea achizițiilor la mai mulți indivizi. Pionierul în dezvoltarea acestui tip de scale a fost Arnold Gesell, psiholog și pediatru care a publicat Gesell Developmental Schedules în 1940 pe baza unei serii de studii longitudinale conduse de el și colaboratorii săi de la Yale de-a lungul unei perioade de câteva decenii (Ames, 1989).

The Provence Birth-to-Three Developmental Profile. Un exemplu curent de instrument care utilizează scale la nivel ordinal este Provence Birth-to-Three Developmental Profile („Profilul Provence"), care este o parte din Infant-Toddler Developmental Assessment (IDA; Provence, Erikson, Vater, & Palmieri, 1995). IDA este un sistem integrat construit pentru a ajuta la identificarea timpurie a copiilor care, din punctul de vedere al dezvoltării, prezintă un risc și, posibil, au nevoie de monitorizare sau de intervenție. Prin observație în mediul natural și rapoarte parentale, Profilul Provence oferă informații despre momentul de timp în care un copil atinge anumite puncte de cotitură în dezvoltare în opt domenii, în relație cu vârsta cronologică a copilului. Domeniile de dezvoltare sunt Comportament motoriu grosier, Comportament motoriu fin, Relații cu obiecte neanimate, Limbaj/Comunicare, Autoajutorare, Relații cu persoanele, Emoții și stări

afective (Afecte) și Comportament de adaptare. Pentru fiecare dintre aceste domenii, profilul grupează itemii în intervale de vârstă de la 0 la 42 de luni. Intervalele de vârstă sunt de 2 luni la vârstele cele mai mici și de 18 luni pentru anumite domenii, la vârste mai mari. Cele mai multe se întind între 3 și 6 luni.

Tabelul 3.1 Exemplu de itemi din Profilul Provence pentru evaluarea dezvoltării bebelușului-copilului mic.

Domeniu	Intervalul de vârstă (în luni)	Item
Comportament motoriu grosier	4 la 7	Stă singur scurt timp
	7 la 10	Se agață pentru a se ridica
	13 la 18	Merge bine singur
	30 la 36	Urcă și coboară trepte
Limbaj/comunicare	4 la 7	Râde tare
	7 la 10	Răspunde la „nu"
	13 la 18	Arată pantofii atunci când i se cere
	30 la 36	Știe versuri sau cântece
Autoajutorare	4 la 7	Recuperează păpușa sau sticla pierdută
	7 la 10	Îndepărtează mâna adultului
	13 la 18	În parte se hrănește singur cu lingura sau cu degetele
	30 la 36	Se încalță cu pantofii

Sursa: Adaptat după *Infant-Toddler Developmental Assessment (IDA) Administration Manual* de Sally Provence, Joanna Erikson, Susan Vater și Saro Palmeri și reprodus cu permisiunea editorului. Copyright © 1995 The Riverside Publishing Company.

Numărul de itemi pentru fiecare grup de vârstă diferă, de asemenea, ca și numărul de itemi care trebuie să fie prezenți sau să poată fi realizați competent pentru a îndeplini criteriul pentru fiecare interval de vârstă. Tabelul 3.1 enumeră patru itemi eșantion pentru fiecare dintre cele trei domenii de dezvoltare ale Profilului Provence din IDA. Scorurile la itemi pentru fiecare interval de vârstă și în fiecare domeniu sunt adăugate pentru a ajunge la o *vârstă a performanței* care poate fi evaluată apoi în comparație cu *vârsta cronologică* a copilului. Discrepanțele dintre nivelele vârstei cronologice și vârstei performanței, dacă există, pot fi apoi utilizate pentru a determina prezența posibilă și măsura întârzierilor în dezvoltarea copilului.

Scale ordinale cu bază teoretică

Scalele ordinale se pot baza pe alți factori decât vârsta cronologică. Câteva teorii, cum ar fi stadiile dezvoltării cognitive din perioada de sugar până la adolescență propuse de Jean Piaget sau teoria dezvoltării simțului moral a lui Lawrence Goldberg, propun o secvență sau o progresie ordonată și invariantă derivată, cel puțin parțial, din observații comportamentale. Unele dintre aceste teorii au generat scale la nivel ordinal construite pentru a evalua nivelul pe care individul l-a atins în cadrul secvenței propuse; aceste instrumente sunt utilizate în primul rând în scopuri legate de cercetare și nu pentru evaluare individuală. Exemplele de astfel de instrumente includ scalele standardizate bazate pe descrierea ordinii în care sunt achiziționate competențele cognitive în perioada copilăriei mici și mari realizate de Piaget, cum ar fi Ordinal Scales of Psychological Development, cunoscută și ca Infant Psychological Development Scales (Užgiris & Hunt, 1975).

Scoruri exprimate în vârste mentale

Noțiunea de scoruri exprimate în vârste mentale a fost discutată în Capitolul 2 în legătură cu raportul IQ al primelor scale de inteligență Stanford-Binet. Scorurile exprimate în vârste mentale derivate din aceste scale au fost calculate pe baza performanței copilului, fiecare câștigând credite în termeni de ani și luni, în funcție de numărul de teste ordonate cronologic care au fost trecute. În lumina dificultăților prezentate de această procedură descrisă în Capitolul 2, acest mod particular de a ajunge la scorurile vârstei mentale a fost abandonat. Totuși, în prezent, câteva teste încă mai oferă norme care sunt prezentate ca *scoruri echivalente cu vârsta* și se bazează pe scorul brut mediu al performanței copiilor din diferite grupuri de vârstă din eșantionul de standardizare. Scorurile echivalente exprimate în vârsta, numite și *vârste date de test*, reprezintă pur și simplu o cale de a stabili o corespondență între performanța unui subiect la un test și performanța medie a grupului de vârstă normativ căruia îi corespunde. De exemplu, dacă scorul brut al unui copil este egal cu scorul brut mediu al copiilor în vârstă de nouă ani din eșantionul normativ, scorul echivalent de vârstă este de 9 ani. În ciuda acestei schimbări în procedurile utilizate pentru a obține scoruri echivalente de vârstă, inegalitățile din rata dezvoltării la diferite vârste rămâne o problemă atunci când este utilizat acest tip de normă de vârstă, pentru că diferențele în achizițiile comportamentale ce pot fi așteptate cu trecerea fiecărui an scad foarte mult din perioada de sugar și copilăria mică și până la adolescență și vârsta adultă. Dacă acest fapt nu este înțeles sau dacă semnificația *vârstei date de test* este extinsă la alte arii decât comportamentul specific eșantionat de test — așa cum se întâmplă, de exemplu, atunci când un adolescent care obține un scor de vârstă la test de 8 ani este descris

ca având „mintea unui copil în vârstă de 8 ani" — utilizarea acestui tip de scoruri poate conduce la concluzii greșite.

Scoruri echivalente cu clasele școlare

Progresia secvențială și uniformitatea relativă a curriculei școlare, mai ales în clasele elementare, oferă baze suplimentare pentru interpretarea scorurilor în termenii normelor de dezvoltare. Astfel, performanța la teste de achiziții de cunoștințe într-un cadru școlar este deseori descrisă la nivelul claselor. Aceste scoruri *exprimate in clase* sunt derivate din localizarea performanței la test în cadrul normelor elevilor la nivelul fiecărei clase — și fracțiuni ale nivelelor de clasă — din eșantionul de standardizare. Dacă spunem, de exemplu, că un copil a obținut scoruri de clasa a șaptea la citire și de clasa a cincea la aritmetică, înseamnă că performanța sa la testul de citire a fost egală cu performanța medie a copiilor din clasa a șaptea din eșantionul de standardizare și că, la testul de aritmetică, performanța sa a fost egală cu cea a celor din clasa a cincea. În ciuda atractivității lor, scorurile de echivalență cu notele pot să ne ghideze greșit din câteva motive. În primul rând, conținutul curriculei și calitatea instruirii variază în diferite școli, sectoare școlare, state și așa mai departe; prin urmare, scorurile exprimate în clase școlare nu oferă un standard uniform. În plus, progresul așteptat în clasele primare ale școlii elementare în termenii achizițiilor academice este mult mai mare decât în gimnaziu sau în liceu; astfel, ca și în cazul unităților de vârstă mentală, o diferență de un an în retard sau accelerare este mult mai semnificativă în clasele primare decât în ultimii ani de școală. Mai mult, dacă un copil care se află în clasa a patra obține la nivelul clasei a șaptea la aritmetică, nu înseamnă că acel copil stăpânește aritmetica de clasa a șaptea; înseamnă că scorul acelui copil la

Nu uitați!

Toate normele de dezvoltare sunt relative, cu excepția faptului că ele reflectă o secvență comportamentală umană sau o progresie *universală*.

- *Scalele la nivel ordinal cu bază teoretică* sunt mai mult sau mai puțin utile, în funcție de teoriile pe care se bazează, dacă acestea sunt întemeiate și pot fi aplicate unui segment de populație sau populației ca întreg.
- *Normele de vârstă mentală sau scalele de scoruri exprimate în vârstă* nu reflectă decât *performanța medie a anumitor grupuri de subiecți aflați la nivele de vârstă specifice* la un moment și loc dat, la un anumit test. Ele sunt supuse schimbării de-a lungul timpului și variază în diferite culturi și subculturi.
- *Normele bazate pe clase sau scale de scoruri echivalente de vârstă* reflectă de asemenea performanța medie a anumitor grupuri de elevi din anumite clase, la un moment și într-un loc dat. De asemenea, ele sunt supuse variației de-a lungul timpului și a curriculei din diferite școli, teritorii școlare și națiuni.

aritmetică este semnificativ deasupra mediei celor de clasa a patra. Mai mult, scorurile exprimate în clase sunt uneori privite în mod eronat drept standarde de performanță pe care toți copiii dintr-o clasă dată trebuie să le îndeplinească, în timp ce ele reprezintă pur și simplu nivele medii de performanță pe care, datorită variabilității inevitabile dintre indivizi, unii elevi le vor îndeplini, alții nu, în timp ce alții le vor depăși.

Normele interne ale grupului

Cele mai multe teste standardizate utilizează un anumit tip de *norme interne ale grupului*. Aceste norme oferă în mod esențial o cale de a evalua performanțele persoanei în comparație cu unul sau mai multe grupuri de referință adecvate. Pentru o interpretare

adecvată a scorurilor obținute la un test cu raportare la normă este necesar să înțelegem procedeele numerice prin care scorurile brute sunt transformate în numeroase tipuri de *scoruri derivate*, utilizate pentru a exprima normele interne ale grupului. Cu toate acestea, este bine să păstrăm în minte faptul că toate diferitele tipuri de scoruri trecute în revistă în această secțiune servesc scopului simplu de a plasa performanța subiectului testat în cadrul unei distribuții normative. Prin urmare, cea mai importantă întrebare cu privire la acest cadru de referință privește componența exactă a grupului sau grupurilor din care sunt obținute normele. Componența eșantionului normativ sau de standardizare este de o importanță deosebită în acest tip de interpretare a scorurilor, deoarece persoanele din acel eșantion stabilesc standardul în raport cu care sunt măsurați ceilalți subiecți cărora li se aplică testul.

Eșantionul normativ

În lumina rolului important jucat de performanța eșantionului normativ, cerința cea mai importantă pentru astfel de eșantioane este ca acestea să fie reprezentative pentru tipul de indivizi cărora li se adresează testul. De exemplu, dacă un test trebuie folosit pentru a evalua abilitățile de citire ale elevilor de școală elementară de la clasa a treia la clasa a cincea din întreaga țară, eșantionul normativ al testului va trebui să reprezinte populația națională de elevi de clasa a treia, a patra și a cincea sub toate aspectele relevante. Componența demografică a populației naționale în raport cu variabile ca gen, etnie, limbă, statut socio-economic, mediu de rezidență urban sau rural, distribuire goegrafică și înscriere în școli publice sau particulare trebuie să fie reflectată în eșantionul normativ al unui astfel de test. În plus, eșantionul

trebuie să fie suficient de mare pentru a asigura stabilitatea valorilor obținute ale performanțelor.

Dimensiunea eșantioanelor normative variază foarte mult în funcție de tipul de test care este standardizat și de ușurința cu care aceste eșantioane pot fi alcătuite. De exemplu, testele de aptitudini aplicate în grup, utilizate în domeniul educațional, pot avea

ATENȚIE!

Deși termenii *eșantion de standardizare*, *eșantion normativ* și *grup de referință* sunt utilizați deseori interșanjabil — aici și în alte părți — și se pot referi de fapt la același grup, strict vorbind, sensul precis al celor trei termeni este oarecum diferit:

- *Eșantionul de standardizare* este grupul de indivizi pe care testul este standardizat inițial în termenii procedurilor de administrare și scorare și de construire a normelor testului. Datele obținute din acest grup sunt prezentate de obicei în manualul care însoțește testul publicat.
- *Eșantionul normativ* este deseori utilizat ca sinonim cu eșantionul de standardizare, dar se poate referi la orice grup din care sunt extrase normele. Normele suplimentare construite după publicarea testului, pentru ca acesta să fie utilizat cu un subgrup distinct, pot apărea în revistele de specialitate sau în manualele tehnice publicate ulterior. Vezi, de exemplu, studiul americanilor vârstnici realizat de Ivnik și colaboratorii săi (1992) la Clinica Mayo, în care datele au fost colectate pentru a oferi norme pentru persoanele aflate dincolo de grupul cu vârsta cea mai înaintată din eșantionul de standardizare al Scalei de inteligență Wechsler pentru adulți — revizuită (WAIS-R).
- *Grupul de referință*, dimpotrivă, este un termen utilizat cu un sens mai larg, pentru a identifica un grup de persoane cu care sunt comparate scorurile obținute la test. Poate fi aplicat grupului de standardizare, eșantionului normativ construit subsecvent, unui grup testat în scopul construirii unor norme locale sau oricărui grup specificat, cum ar fi elevii dintr-o singură clasă sau participanții la o cercetare.

eșantioane normative numărând zeci sau chiar sute de mii de indivizi, în timp ce testele individuale de inteligență, administrate unei singure persoane la un moment dat de către un examinator instruit sunt normate pe eșantioane mult mai mici — de obicei între 1000 și 3000 de indivizi — extrase din populația generală. Testele care necesită eșantioane specializate, cum ar fi membrii unui anumit grup educațional, pot avea eșantioane normative chiar mai mici. Caracterul recent al informației normative este de asemenea important, dacă este necesară evaluarea subiecților testați în raport cu standarde contemporane, așa cum este cazul de obicei.

Factorii relevanți care trebuie luați în considerare în alcătuirea eșantionului normativ variază în funcție de scopul testului și de populația pentru care el va fi utilizat. În cazul unui test construit pentru a detecta deteriorarea congnitivă la vârstnici, de exemplu, variabile ca starea sănătății, situații de trai independent versus instituționalizat și nivelul medicamentației vor fi relevante, alături de variabilele demografice de gen, vârstă, etnie și altele. Info 3.2 enumeră câteva dintre cele mai obișnuite întrebări pe care utilizatorii testului ar trebui să și le pună cu privire la eșantionul normativ atunci când sunt în procesul de evaluare a adecvării unui test la scopurile lor.

Grupurile de referință pot fi definite pe un continuum de lărgime sau specificitate în funcție de tipul de comparații prin care utilizatorii testului trebuie să evalueze scorurile obținute la test. La o extremă, grupul de referință poate fi populația generală a unei întregi națiuni sau chiar o populație multinațională. La celălalt capăt, grupurile de referință pot fi extrase din populații care sunt definite limitat în termeni de statut sau cadru.

Norme de subgrup. Atunci când sunt construite eșantioane mari pentru a reprezenta populații definite larg, normele pot fi raportate

Info 3.2

Informații necesare pentru evaluarea aplicabilității eșantionului normativ

Pentru a evalua adecvarea unui test cu raportare la normă la un scop specific, utilizatorii testului trebuie să aibă cât de multe informații cu privire la eșantionul normativ, inclusiv răspunsuri la întrebări cum ar fi:

- Care este dimensiunea eșantionului normativ?
- Când a fost alcătuit eșantionul?
- Unde a fost alcătuit eșantionul?
- Cum erau indivizii identificați și selectați pentru eșantion?
- Cine a testat eșantionul?
- Care era calificarea examinatorului sau examinatorilor pentru a efectua testarea?
- Care era componența eșantionului normativ în termeni de
  - vârstă?
  - sex?
  - etnie, rasă sau limbă?
  - educație?
  - statut socio-economic?
  - distribuție geografică?
  - orice altă variabilă relevantă care ar putea influența performanța la test, cum ar fi starea sănătății fizice și mentale sau apartenența la un grup atipic?

Utilizatorii testului pot evalua adecvarea unui test cu raportare la normă la scopurile lor specifice doar atunci când răspunsurile la aceste întrebări sunt oferite în manualul testului sau în documentația testului.

cumulat sau pot fi separate în *norme de subgrup.* Dat fiind că au o dimensiune suficient de mare — și sunt reprezentative pentru categoriile lor — pot fi formate subgrupuri în termeni de vârstă, sex, ocupație, etnie, nivel educațional sau orice altă variabilă care

ar putea avea un impact semnificativ asupra scorurilor obținute la test sau ar putea duce la comparații ce prezintă interes. Normele de subgrup pot fi obținute, de asemenea, după ce un test a fost standardizat și publicat, pentru a suplimenta și extinde aplicabilitatea testului. De exemplu, înainte ca MMPI să fie revizuit pentru a se crea MMPI-2 și o formă separată pentru adolescenți (MMPI — A), utilizatorii testului original — care a fost normat exclusiv pe adulți — au construit norme speciale de subgrup pentru adolescenți la diferite nivele de vârstă (vezi Archer, 1987).

Norme locale. Pe de altă parte, există anumite situații în care utilizatorii testului pot dori să evalueze scoruri pe baza grupurilor de referință extrase dintr-un cadru geografic sau instituțional specific. În astfel de cazuri, utilizatorii testului pot alege să construiască un set de *norme locale* pentru membrii unei populații mai strict definite cum ar fi angajații unei anumite companii sau studenții unei anumite universități. Normele locale pot fi utilizate pentru evaluarea performanțelor studenților sau angajaților dintr-un cadru dat sau pentru luarea deciziilor cu privire la candidații la o școală sau la un loc de muncă în raport cu standardele acelui loc de muncă sau ale acelei instituții.

Norme convenabile. Ocazional, din motive de promptitudine sau constrângeri financiare, cei care construiesc testele utilizează norme bazate pe un grup de oameni care se întâmplă pur și simplu să fie disponibili în momentul în care este construit testul. Aceste *norme convenabile* au o utilitate limitată pentru că nu sunt reprezentative pentru nicio populație definită; ele sunt deseori compuse din indivizi care sunt ușor accesibili celor care construiesc testele, cum ar fi elevii dintr-o clasă de liceu sau rezidenții dintr-un anumit azil de bătrâni. În astfel de cazuri, natura eșantionului normativ trebuie clar specificată potențialilor utilizatori ai testului.

Percentile

Scorurile rangului percentilic, discutate deja în Capitolul 2, sunt metodele cele mai directe și mai frecvent utilizate pentru a transmite rezultatele unui test cu raportare la normă. Avantajele lor principale sunt că ele pot fi înțelese cu ușurință de persoanele testate și pot fi aplicate celor mai multe tipuri de teste și de populații de testare. Un *scor exprimat în percentile* indică poziția relativă a unui subiect individual căruia i s-a aplicat testul comparativ cu un grup de referință cum ar fi eșantionul de standardizare; în mod specific, reprezintă procentul de persoane din grupul de referință care au obținut scoruri egale sau mai mici decât un scor brut dat. Astfel, cel mai mare scor percentilă indică

ATENȚIE!

Datorită similarității dintre doi termeni, scorurile exprimate în percentile sunt deseori confundate cu scorurile exprimate în procente. Aceste două tipuri de scoruri sunt, de fapt, destul de diferite și utilizează cadre de referință total diferite:

- *Percentilele* sunt scoruri care reflectă rangul sau poziția performanței unui individ la un test în comparație cu un grup de referință; cadrul de referință al acestor scoruri este dat de ceilalți oameni.
- *Scorurile exprimate în procente* reflectă numărul de răspunsuri corecte pe care un individ le-a obținut din numărul total de răspunsuri corecte la un test; cadrul lor de referință este conținutul întregului test.

O cale de a evita confuzia este de a ne face un obicei din a folosi simbolul procent (%) doar pentru scorurile exprimate în procente și de a utiliza o abreviere diferită, cum ar fi *PR* sau *%'ilă*, pentru a nota scorurile exprimate în percentile.

cele mai mari scoruri brute la dimensiunea măsurată de test; a 50-a percentilă (P_{50}) sau mediana corespunde punctului scorului brut care separă jumătatea superioară de jumătatea inferioară a distribuției scorurilor grupului de referință. Într-o distribuție normală, a 50-a percentilă reprezintă, de asemenea, nivelul mediu de performanță al grupului.

Un avantaj suplimentar al scorurilor rangului percentilic intră în joc atunci când există mai mult de un grup normativ pentru același test sau când grupurile normative sunt subdivizate în categorii, cum ar fi genul, vârsta sau etnia. Atunci când sunt disponibile norme suplimentare, un scor brut poate fi localizat în cadrul distribuțiilor a două sau mai multe grupuri sau subgrupuri diferite și poate fi convertit cu ușurință în rang percentilic. De exemplu, scorurile unui inventar de interese pentru diferite grupuri ocupaționale sunt deseori raportate pentru bărbați și femei ca norme de grup separate, astfel încât persoanele testate pot vedea poziția lor pentru un interes dat și pe o scală ocupațională comparativ cu ambele grupuri. Această informație este deosebit de utilă pentru cei care au în vedere o ocupație care este semnificativ segregată în funcție de sex, cum ar fi cea de inginer sau de asistent medical. Separarea normelor permite indivizilor să determine forța relativă a intereselor exprimate în comparație cu membrii grupurilor de ambele sexe.

Dacă scorurile obținute la teste au fost distribuite echilibrat în interval, vor conduce la un poligon de frecvențe rectagular, iar percentilele vor fi probabil scorurile recomandate în aproape toate situațiile. Cu toate acestea, așa cum se poate vedea în Figura 2.2 și, mai târziu în acest capitol în Figura 3.1, într-o distribuție normală majoritatea scorurilor tind să se grupeze în jurul unei valori centrale și se dispersează mai mult la extreme. Acest fapt, care

este adevărat și pentru multe distribuții ne-normale, înseamnă că unitățile de scor percentilă sunt marcate de obicei inegal, în diferite puncte cuprinse în amplitudinea scorurilor. Într-o distribuție normală sau aproape normală cum sunt cele obținute în cazul celor mai multe teste, procentul de oameni care obțin un scor apropiat de mijlocul distribuției este mai mare decât procentul celor care obțin scoruri apropiate de extreme. Prin urmare, orice diferență dată în unități de scor al rangului percentilic amplifică diferența evidentă a performanței relative a indivizilor ale căror scoruri se află în zona medie și comprimă extinderea aparentă a diferenței dintre performanțele relative ale indivizilor situați la capetele extreme ale distribuțiilor.

Un alt dezavantaj al scorurilor exprimate în rang percentilic se referă la cele mai extreme scoruri ale distribuției. Cu siguranță, pentru orice grup normativ dat există întotdeauna cel mai înalt scor și cel mai scăzut scor. Atât timp cât scorurile sunt interpretate strict în raport cu un eșantion normativ specific, se poate spune că cel mai mare scor se află la percentila 100, deoarece toate cazurile se află la nivelul acelui scor sau sub el; tehnic, putem chiar descrie scorul aflat sub scorul cel mai scăzut obținut de o persoană din eșantionul specific ca fiind scorul dat de percentila zero, deși aceasta nu se face de obicei. Totuși, în termenii populației mai mari reprezentate de eșantionul normativ, interpretarea acestor scoruri este problematică.

Palierul superior și inferior al testului. Problema pusă de indivizii aflați la capetele extreme ale spectrului de aptitudini pentru care este construit un test are cea mai mare relevanță în contextul construirii testului, discutat în Capitolul 6. Cu toate acestea, în acest punct merită să notăm faptul că indivizii implicați în standardizarea unui test stabilesc limitele superioare și inferioare ale

Aplicație practică

Utilizarea scorurilor de rang percentilic
Dezavantajele unităților inegale

Subtestul Vocabular al Scalei de Inteligență Wechsler pentru Adulți — ediția a 3-a (WAIS-III), un test pentru indivizii cu vârsta cuprinsă între 16 și 89 de ani, constă din 33 de cuvinte. Fiecare definiție a unui cuvânt poate primi un scor de 0, 1 sau 2 puncte. Astfel, scorurile brute la subtest pot fi cuprinse între 0 și 66, în funcție de calitatea răspunsurilor.

Manualul WAIS-III (Wechsler, 1997) prezintă performanța eșantioanelor de standardizare în tabele pentru diferite intervale de vârstă. Pentru indivizii cu vârste între 45 și 54 de ani, tabelul arată că scorurile brute cuprinse între 45 și 48 de puncte au rangul percentilic 50, iar scorurile brute cuprinse între 49 și 51 de puncte se află la percentila 63. Prin contrast, scorurile brute dintre 15 și 19 puncte au rangul percentilic 2 și cele dintre 20 și 24 de puncte au rangul percentilic 5.

Aceasta subliniază în mod clar problema inegalității unităților scorurilor exprimate în percentile: pentru persoanele din grupul de vârstă de la 45 la 54 de ani, o diferență de 6 puncte (de la 45 la 51) în mijlocul distribuției scorurilor brute duce la o diferență de 13 unități ale rangului percentilic (de la a 50-a la a 63-a percentilă), în timp ce o diferență a scorurilor brute de 9 puncte (de la 15 la 24) la capătul inferior al intervalului duce la o diferență de doar 3 unități ale rangului percentilic (de la a 2-a la a 5-a percentilă)

performanței la acel test. Dacă o persoană testată atinge cel mai mare scor ce poate fi obținut la un test deja standardizat, înseamnă că *plafonul testului*, sau nivelul maxim de dificultate al testului, este insuficient: nu se poate ști cât de mare ar fi putut fi scorul obținut de acea persoană dacă în test ar fi existat itemi suplimentari sau itemi cu o dificultate mai mare. În mod similar, dacă o persoană greșește toți itemii prezentați într-un test sau obține

un scor mai mic decât al tuturor persoanelor din eșantionul normativ, problema este aceea a unui *prag inferior* insuficient. În astfel de cazuri, indivizii în chestiune nu au fost testați în mod adecvat.

Scoruri standard

O cale de a depăși problema inegalității unităților percentile și de a transmite încă semnificația scorurilor testului în raport cu un grup normativ sau un grup de referință este de a transforma scorurile brute pe scale care exprimă poziția scorurilor în raport cu media, în unități de abatere standard. Aceasta se poate realiza prin intermediul unor transformări liniare simple. O *transformare liniară* modifică unitățile în care sunt exprimate scorurile, lăsând nemodificate interrelațiile dintre acestea. Cu alte cuvinte, forma unei distribuții a scalei scorurilor derivată liniar pentru un grup dat este aceeași cu cea a distribuției inițiale a scorurilor brute. Un mare avantaj al acestei metode este faptul că scorurile testelor cu o distribuție normală care au medii, abateri standard și intervale diferite pot fi comparate — și mediate — din momentul în care au fost transformate liniar în scoruri ale unei scale comune, atâta timp cât este utilizat același grup de referință.

Prima transformare liniară realizată asupra scorurilor brute este de a le transforma în abateri standard ale scorurilor sau

ATENȚIE!

Ca și în capitolul anterior, formulele și procedeele statistice prezentate aici sunt cele *esențiale*, necesare pentru o înțelegere bazală a transformării scorurilor. Deși operațiile statistice descrise în această carte pot fi realizate cu programe pe calculator cum ar fi SPSS și SAS, și, de obicei, așa se întâmplă, formulele și pașii pe care îi presupune această transformare trebuie înțelese pentru a înțelege aspectele esențiale ale semnificației diferitelor scoruri.

scoruri z. Un *scor z* (vezi Anexa C) exprimă distanța dintre un scor brut și media grupului de referință în termenii abaterii standard a grupului de referință. Să ne reamintim că media și abaterea standard a scorurilor z sunt zero și, respectiv, 1, și că distribuția scorurilor z este simetrică bilateral, cu o jumătate din cazuri de fiecare parte a mediei. Poziția scorurilor z în raport cu media este indicată de utilizarea unui semn pozitiv (sau niciun semn) pentru scorurile z care se află deasupra mediei și un semn negativ pentru cele aflate sub medie. Astfel, semnul unui scor z indică direcția în care un scor deviază de la media unui grup, în timp ce valoarea sa numerică reflectă distanța scorului față de medie în unități de abatere standard (SD). De exemplu, un scor z de + 1,25 indică faptul că scorul brut inițial este la 1 1/4 unități *SD deasupra* mediei grupului, în timp ce un scor brut care se află la 3/4 unități *SD sub* medie se transformă într-un scor z de — 0,75. Dacă distribuția scorurilor pentru eșantionul de referință este normală, scorurile z pot fi transformate imediat în percentile prin consultarea Tabelului ariilor de sub curba normală prezentat și explicat în Anexa C.

Info 3.3 prezintă formulele de transformare liniară utilizate pentru a deriva scorurile z din scoruri brute și pentru a transforma scorurile z în alte tipuri de scoruri standard. Deoarece transformarea scorurilor brute în scoruri z este de obicei prima în procesul de transformare a scorurilor, scorurile z sunt considerate a fi exemplul principal de scoruri standard și deseori sunt denumite pur și simplu scoruri standard. Aceasta le distinge de alte tipuri familiare de scoruri standard sau scoruri derivate, cum este IQ, care a ajuns să fie asociat cu teste specifice și către care ne vom îndrepta atenția în continuare.

Sisteme suplimentare pentru obținerea scorurilor standard. Deși scorurile z ne permit să aflăm imediat mărimea și direcția diferenței

Info 3.3

Formula pentru transformarea scorurilor brute în scoruri z:

$$z = \frac{X - \bar{X}}{SD_X}$$

unde X = Scor brut
$\bar{X}$ = Media grupului de referință
SD_X = Abaterea standard (*SD*) a grupului de referință

Formula de transformare a scorurilor z în alte scoruri standard derivate:

Scor standard nou = (scor z) (noua SD)+ noua medie

Exemplu: Pentru a transforma un scor z de +1,00 într-un scor IQ cu $M = 100$ și $SD = 15$

scor IQ = (+1,00) (15) + 100 = 115

dintre orice scor și media distribuției din care face parte, ele presupun valori negative și zecimale. Din acest motiv, scorurile z sunt supuse de obicei unei transformări liniare suplimentare. Scopul sistemelor de scoruri standard care rezultă din aceste transformări consecutive este pur și simplu acela de a exprima rezultatele obținute la test într-o formă mai convenabilă.

Numerele alese ca medii și abateri standard în transformarea scorurilor z în diferite alte formate de scoruri standard sunt arbitrare. Cu toate acestea, prin utilizarea lor frecventă în diferite contexte, aceste formate de scoruri au devenit familiare și sunt înțelese într-un anumit mod — ceea ce nu poate fi garantat întotdeauna — de către cei care le utilizează. Figura 3.1 prezintă curba normală cu abscise pentru percentile, scoruri z și următoarele sisteme de scoruri standard larg utilizate:

- *Scoruri-T* ($M = 50$, $SD = 10$), utilizate în multe inventare de personalitate cum sunt Inventarul Multifazic de Personalitate Minnesota (MMPI) și Inventarul Psihologic California (CPI).
- *Scoruri College Entrance Examination Board (CEEB)* ($M = 500$, $SD = 100$), utilizate de SAT și de Educational Testing Service pentru mulți dintre absolvenții lor și pentru programele de testare la admiterea în școli profesionale, cum ar fi Graduate Record Exam (GRE).
- *Scoruri ale subtestelor scalei Wechsler* ($M = 10$, $SD = 3$), utilizate pentru toate subtestele scalelor Wechsler, ca și pentru subtestele câtorva alte instrumente.
- *IQ-ul pe scala Wechsler* ($M = 100$, $SD = 15$), utilizat pentru a sintetiza scorurile la toate scalele Wechster și la alte teste, incluzând multe instrumente care nu-și etichetează scorurile ca „IQ".
- *Otis-Lennon School Ability Indices* ($M = 100$, $SD = 16$), utilizat în Otis-Lennon School Ability Test (OLSAT), care este titlul actual al testelor cu administrare în grup ce au apărut odată cu Otis Group Intelligence Scale.

Indicatorii OLSAT sunt incluși în Figura 3.1, ca un exemplu de sistem de scoruri standard cu o medie de 100 și o abatere standard diferită de 15, pentru a ilustra arbitrariul alegerii unităților în sistemele de scoruri standard. Deși cele mai multe sisteme de scoruri standard utilizează o medie de 100 selectează 15 ca abatere standard, există alte sisteme care utilizează alte abateri standard, cum ar fi 12 sau 18. Aceste alegeri alternative pot duce la diferențe semnificative în interpretarea scorurilor. De exemplu, dacă două teste cu o distribuție normală — ambele cu

media de 100 — au abateri standard de 12 și 15, un scor standard de 112 la primul test se va transforma într-un scor z de +1,00 și rangul său va fi dat de a 84-a percentilă, în timp ce la al doilea test

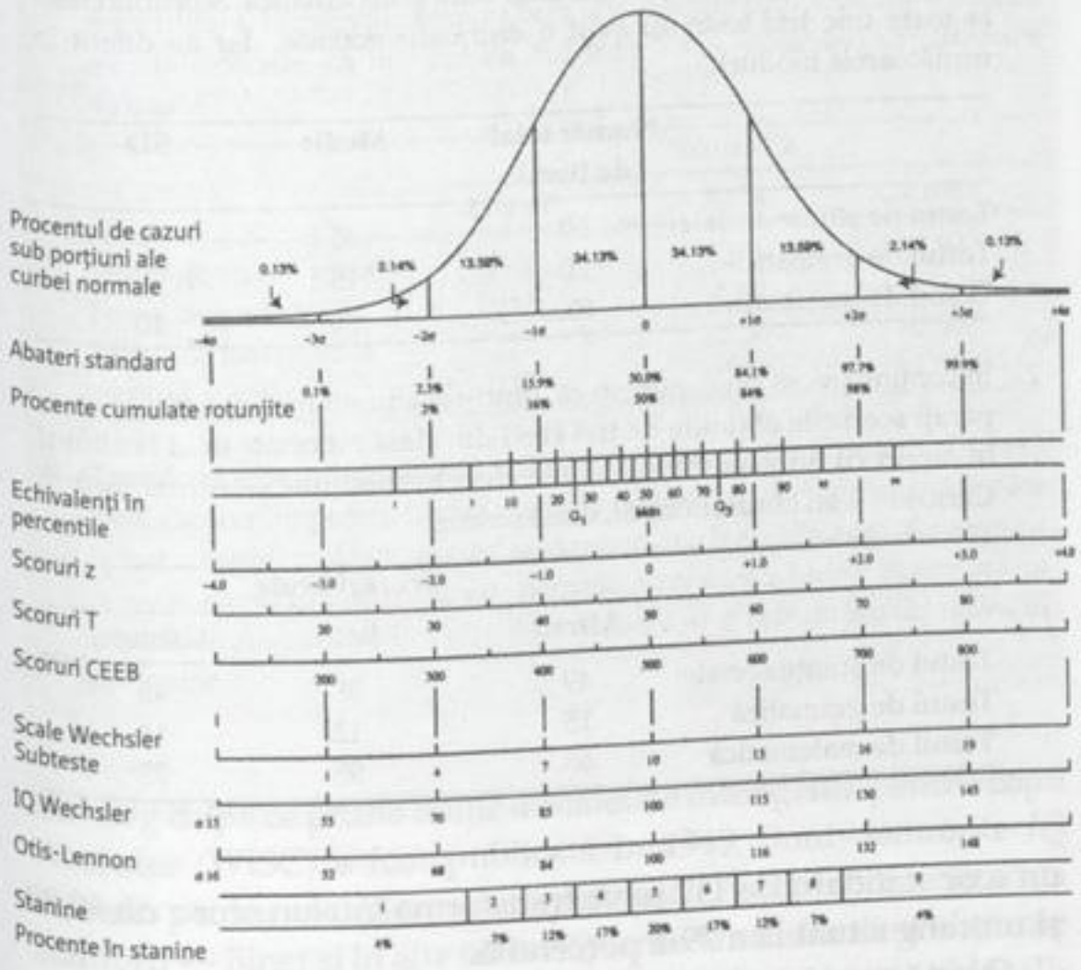

Figura 3.1 Curba normală, percentile și scoruri standard selecționate

Notă: Adaptată după *Test Service Notebook #148* al The Psychological Corporation.

Aplicație practică

Transformarea scorurilor brute în scoruri *z*

1. Să presupunem că toți elevii dintr-o clasă a opta au dat testul de achiziții de cunoștințe din științe sociale, gramatică și matematică. Scorurile clasei la toate cele trei teste au avut o *distribuție normală*, dar au diferit în următoarele moduri:

	Număr total de itemi	Medie	SD
Testul de științe sociale	50	35	5
Testul de gramatică	20	15	3
Testul de matematică	100	70	10

2. În continuare, să presupunem că, dintr-un anumit motiv, doriți să comparați scorurile obținute de trei elevi din clasă raportate unul la altul și în raport cu întreaga clasă. Cei trei elevi în chestiune — Alfred, Beth și Carmen — au obținut următoarele scoruri:

	Scoruri brute		
	Alfred	Beth	Carmen
Testul de științe sociale	49	38	48
Testul de gramatică	15	12	18
Testul de matematică	68	95	75

un scor standard de 112 se va transforma într-un scor *z* de +0,80 și un rang situat la a 79-a percentilă.

O notă cu privire la scorurile IQ pe scala Wechsler. Scorurile cunoscute drept *IQ Wechsler* au fost introduse de David Wechsler în 1939 pentru a fi utilizate în prima sa scală de inteligență Wechsler-Bellevue I, care a devenit mai târziu Scala de inteligență Wechsler pentru adulți (WAIS). Aceste scoruri au fost utilizate

3. Aceste scoruri nu pot fi comparate și nu se poate face o medie între teste, deoarece sunt scoruri pe scale diferite. Pentru a compara scorurile, chiar și la un singur test, trebuie să ne raportăm la media, abaterea standard (SD) și numărul de itemi pentru fiecare test. O cale ușoară de a le compara este de a converti fiecare scor brut în scoruri *z* — extrăgând media respectivului test din fiecare scor brut și împărțind prin abaterile standard corespunzătoare, ca în formula din Info 3.3 — obținând următoarele rezultate:

	Scoruri z		
	Alfred	Beth	Carmen
Testul de științe sociale	+2,80	+0,60	+2,60
Testul de gramatică	0,00	-1,00	+1,00
Testul de matematică	-0,20	+2,50	+0,50
rezultat mediat	+0,87	+0,70	+1,37

4. Transformările liniare ale scorurilor brute în scoruri *z* ne permit să facem media scorurilor pentru fiecare elev și să comparăm performanțele elevilor la toate testele unul față de altul și în raport cu clasa. *Mai mult, din moment ce am presupus că distribuția este normală pentru toate testele, putem utiliza tabelul din Anexa C pentru a traduce fiecare scor z într-un scor al rangului percentilic.*

mai larg după ce prima ediție a Scalei de inteligență pentru copii Wechsler (WISC) a fost publicată în 1949. Sunt denumite *IQ Wechsler* pentru a le diferenția de *raportul IQ* inițial utilizat în scala Stanford — Binet și în alte scale. IQ-urile Wechsler sunt obținute prin însumarea scorurilor obținute la diferite subteste și localizarea acestei sume în tabloul normativ adecvat și nu prin formula (Vârsta mentală/Vârstă cronologică) x 100.

Scala de scoruri de tip IQ Wechsler a fost adoptată de mulți dintre cei care construiesc teste pentru a exprima scorurile sintetice

ale diferitelor teste, inclusiv de cea mai recentă ediție a scalei Stanford-Binet care folosește o unitate de abatere standard de 15 în locul abaterii sale standard inițiale de 16. În numeroasele teste care utilizează scorurile standard cu media de 100 și abaterea standard de 15 se întâlnesc toate instrumentele majore destinate să evalueze funcționarea cognitivă generală, cum ar fi seriile de teste Kaufman (e.g., Kaufman-Assessment Battery for Children, Kaufman Adolescent and Adult Intelligence Test, etc.), Differential Ability Scales, Das-Naglieri Cognitive Assessment System și multe altele. Deși au același tip de unități ca testele Wechsler pentru scorurile globale sau sintetice, toate aceste teste mai noi au renunțat la utilizarea termenului *IQ* pentru denumirea scorurilor. Aceasta este o schimbare prudentă, deoarece IQ-urile Wechsler *nu* sunt rapoarte. Mai mult, așa cum se poate vedea din titlurile unora dintre instrumentele construite mai recent, autorii testelor renunță din ce în ce mai mult la utilizarea cuvântului *inteligență* pentru denumirea constructului determinat prin testele lor în favoarea unor alți termeni, mai neutri.

Transformări nonliniare

Nu toate transformările scorurilor sunt liniare. *Transformările nonliniare* sunt cele care convertesc o distribuție de scoruri brute într-o distribuție cu o formă diferită de cea inițială. Aceasta se poate face utilizând în construirea testului metode care permit o flexibilitate mai mare în manevrarea distribuțiilor scorurilor brute decât conversiile liniare. Deși anumite transformări nonliniare ale scorurilor presupun operații complexe care depășesc domeniul acestei cărți, altele nu presupun acest lucru. De exemplu, transformarea scorurilor brute cu distribuție normală în scoruri de rang percentilic, pe care am luat-o deja în considerare, este o

conversie nonliniară. Este realizată prin transformarea fiecărui scor brut într-un scor z și localizarea scorului z în Tabelul ariilor situate sub curba normală pentru a obține proporția sau procentul ariei curbei normale care se află sub acel punct. În această secțiune, vom discuta două alte tipuri de scoruri standard, larg răspândite, derivate nonliniar.

Scorurile standard normalizate sunt utilizate atunci când o distribuție de scoruri aproximează distribuția normală, dar nu se suprapune exact cu aceasta. Pentru a normaliza scorurile trebuie să găsim mai întâi procentul de persoane din eșantionul de referință care se află la nivelul fiecărui scor brut sau sub el (vezi, e.g. coloana Procent cumulat (PC) pentru distribuția celor 60 de scoruri la test din Tabelul 2.3). Apoi, procentele sunt transformate în proporții. Aceste proporții sunt apoi localizate în Tabelul ariilor situate sub curba normală pentru a obține scoruri *z* corespunzătoare acelor arii (vezi Anexa C). Scorurile standard obținute în această manieră nu pot fi distinse ca formă de cele obținute prin formula transformării liniare, dar întotdeauna trebuie identificate ca scoruri standard normalizate pentru a atenționa utilizatorul testului că ele provin dintr-o distribuție care nu a fost normală. În momentul în care sunt obținute scorurile standard normalizate, acestea pot fi transformate în alte sisteme de scoruri standard convenabile pe care le-am discutat, cum ar fi scoruri T, IQ Wechsler sau scoruri CEEB, utilizând aceeași procedură utilizată pentru scorurile *z* derivate liniar descrisă în Info 2.3.

Staninele au fost inițial introduse de U.S. Air Force în timpul celui de-al Doilea Război Mondial. Scalele *„stan*dard *nine"* [standard nouă] sau *stanine* transformă toate scorurile dintr-o distribuție în numere alcătuite dintr-o cifră de la 1 la 9. Acest instrument are avantajul distinct de a reduce timpul și efortul necesar pentru

Aplicație practică

Transformarea scorurilor z în diferite scoruri standard

1. Pentru a ilustra conversia scorurilor z în diferite tipuri de sisteme de scor standard, să ne întoarcem la exemplul anterior al celor trei elevi dintr-o clasă a opta ale căror scoruri z la trei teste de achiziții de cunoștințe cu distribuție normală au fost:

	Scoruri z		
	Alfred	**Beth**	**Carmen**
Testul de științe sociale	+2,80	+0,60	+2,60
Testul de gramatică	0,00	-1,00	+1,00
Testul de matematică	-0,20	+2,50	+0,50

2. Pentru a transforma aceste scoruri z în scoruri pe o scală mai convenabilă, aplicăm formula de transformare în alte scoruri standard, prezentată mai sus în Info 3.3.

Scor standard nou = (scor z) (noua SD)+ noua medie

a introduce scorurile în computer pentru a fi stocate și procesate ulterior. Transformările în stanine utilizează de asemenea distribuțiile de frecvențe cumulate și procente cumulate, cum sunt cele din Tabelul 2.3; scorurile stanine sunt atribuite pe baza procentului de cazuri care corespunde intervalului unei stanine date. Tabelul 3.2, Partea A, arată procentele curbei normale din cadrul fiecărei unități stanine și procentele cumulate utilizate în transformarea scorurilor brute în stanine. Partea B conține câteva exemple de transformări în stanine, utilizând unele dintre scorurile din Tabelul 2.3. Figura 3.1 ilustrează poziția scorurilor stanine pe curba normală. Așa cum se poate vedea din figură, stanina 5 cuprinde

3. Utilizând mediile și abaterile standard adecvate pentru fiecare sistem, obținem următoarele scoruri:

	Alfred	**Beth**	**Carmen**
Scoruri T (M = 50, SD = 10)			
Testul de științe sociale	78	56	76
Testul de gramatică	50	40	60
Testul de matematică	48	75	55
Scoruri de tip CEEB (M = 500, SD = 100)			
Testul de științe sociale	780	560	760
Testul de gramatică	500	400	600
Testul de matematică	480	750	550
IQ Scală Wechsler (M = 100, SD = 15)			
Testul de științe sociale	142	109	139
Testul de gramatică	100	85	115
Testul de matematică	97	138	108

4. Observați că, deoarece toate acestea au fost transformări liniare, *scorurile se află exact în aceeași poziție unul în raport cu celălalt, indiferent de diferențele dintre unități.* Acest fapt poate fi observat și mai clar în relația dintre scorurile T și scorurile CEEB, în care media și abaterea standard pentru CEEB este de 10 ori mai mare decât media și abaterea standard a scorurilor T; prin urmare, în exemplul nostru, fiecare scor CEEB este egal cu scorul T corespunzător înmulțit cu 10.

20% din scorurile din zona de mijloc. Media scalei în stanine este 5, și abaterea standard este aproximativ 2. Chiar dacă scala în stanine este economică și simplă, concizia și simplitatea sa duc și la o anumită pierdere a preciziei.

Aplicație practică

Scoruri standard normalizate

Pentru a demonstra procesul de normalizare a scorurilor unei distribuții care nu se conformează curbei normale, ci o aproximează, putem folosi cinci dintre scorurile brute ale distribuției de 60 de scoruri la test prezentate în Tabelul 2.3 în Capitolul 2. Scorurile brute selectate arbitrar pentru acest exercițiu sunt 49, 40, 39, 36 și 29.

Pașii acestei conversii sunt următorii:

Scor brut → procent cumulat (PC) → proporție cumulată (pc) → Scor *z* normalizat

1. Scorul brut și procentul cumulat sunt localizate în distribuție.
2. Proporția cumulată este procentul cumulat împărțit la 100.
3. *Un scor z normalizat* este obținut din Tabelul ariilor situate sub curba normală din Anexa C prin descoperirea proporției ariei de sub curbă care este cea mai apropiată de proporția cumulată a unui scor dat. Pentru scorurile cu proporții cumulate peste 0,50, trebuie utilizate ariile din porțiunile mai mari — din coloana (3) a tabelului — pentru a obține scoruri *z* normalizate care vor avea un semn pozitiv. Pentru scorurile cu proporții cumulate sub 0,50, trebuie utilizate ariile din porțiunea mai mică — din coloana (4) a tabelului — iar scorurile *z* normalizate rezultate vor avea semne negative.
4. Aceste procedee duc la următoarele rezultate:

Scor brut	*% cumulat*	*pc*	Scor z normalizat[a]
49	98,3	0,983	+2,12
40	53,3	0,533	+0,08
39	45,0	0,450	-0,13
36	23,3	0,233	-0,73
29	1,7	0,017	-2,12

[a]Cu privire la distribuțiile lor, în termenii poziției scorurilor brute inițiale pe care le reprezintă, scorurile standard normalizate au aceeași semnificație ca și scorurile standard derivate liniar. Ele pot fi transformate în diferite alte sisteme de scoruri standard, dar trebuie identificate întotdeauna ca fiind derivate nonliniar printr-o procedură de normalizare.

Tabelul 3.2 Transformarea scorurilor brute în scoruri stanine

A. Procentele curbei normale utilizate în transformarea în stanine

	Stanine								
	1	2	3	4	5	6	7	8	9
Procentul de cazuri din fiecare stanină	4	7	12	17	20	17	12	7	4
Procentul cumulat pentru fiecare stanină	4	11	23	40	60	77	89	96	100

B. Scoruri la test selectate din Tabelul 2.3 convertite în stanine

Scoruri selectate[a]	% cumulat	Scoruri în stanine
49	98,3	9
47	93,3	8
45	83,3	7
43	71,7	6
41	60,0	5
39	45,0	5
37	30,0	4
35	20,0	3
33	11,7	3
31	5,0	2
29	1,7	1

[a]Vezi Tabelul 2.3 din Capitolul 2 pentru distribuția completă a celor 60 de scoruri la test.

Comparații între teste

În cele mai multe situații, scorurile obținute la un test cu raportare la normă nu pot fi comparate în afară de cazul în care sunt obținute de la același test, utilizând aceeași distribuție normativă. Un motiv suplimentar pentru incapacitatea de a compara scorurile obținute la test apare din diferențele dintre unitățile scalelor, precum diferitele dimensiuni ale unităților SD discutate mai sus în legătură cu IQ-ul Wechsler. Mai mult, chiar și atunci când testele, normele și unitățile scalei utilizate sunt aceleași, scorurile obținute la test nu au în mod necesar aceeași semnificație.

Atunci când scorurile obținute la test sunt utilizate în contextul evaluării individuale, trebuie păstrat în minte faptul că numeroși alți factori externi testului pot intra în rezultatele la test (ex.: motivația și formarea subiectului, influența examinatorilor, circumstanțele în care este aplicat testul).

Tehnici de echivalare

În ciuda avertismentelor subliniate în secțiunea precedentă, există un număr de situații în care este necesar sau dezirabil să comparăm scorurile indivizilor sau grupurilor în timp sau în diferite situații de funcționare psihică, în raport cu o normă uniformă. Pentru astfel de situații, cei care construiesc și publică teste au inventat câteva moduri de a obține o *anumită* comparabilitate a scorurilor dintre teste. Esențial, toate acestea sunt destinate plasării scorurilor obținute la test într-un cadru comun de referință. Un beneficiu suplimentar al tehnicilor de echivalare este că ele ar putea reduce considerabil costurile și timpul necesare pentru standardizarea testelor. Cele mai multe proceduri de echivalare presupun detalii foarte tehnice, descrise în monografia lui Angoff (1984) și în alte surse (Petersen et al., 1989). Următoarele descrieri concise ale unora dintre cel mai frecvent utilizate abordări îi pot oferi cititorului explicații de bază despre acestea.

Nu uitați!

1. Compararea scorurilor obținute la un test nu are sens dacă:
   - Testele sau versiunile testului sunt diferite
   - Grupurile de referință sunt diferite sau
   - Scalele scorurilor diferă, cu excepția situației în care testele, grupurile sau scalele au fost echivalate în acest scop. Monografia lui Angoff, *Scales, norms, and equivalent scores,* este una dintre cele mai bune surse de informare despre procedurile de echivalare.
2. Chiar și atunci când scorurile obținute la test pot fi comparate, contextul în care are loc testarea, precum și educația, experiența și circumstanțele sociale ale persoanelor testate trebuie să fie luate în calcul în interpretarea scorurilor la test.

- *Formele alternative* constau din două sau mai multe versiuni ale unui test menite a fi utilizate interșanjabil, care urmăresc același scop și sunt administrate în manieră identică. Crearea formelor alternative care sunt asemănătoare în conținutul pe care-l acoperă, dar variază în itemii specifici pe care îi conțin este una dintre cele mai simple modalități de a produce teste comparabile. O modalitate mai strictă de a obține teste comparabile poate fi atinsă cu tipul de versiuni alternative ale testului, cunoscute drept *forme paralele.* Aceste forme sunt echivalente nu numai în privința acoperirii conținutului și a procedurilor, ci și a unor caracteristici statistice precum ar fi mediile și abaterile standard ale scorurilor brute, precum și a unor indicatori ai fidelității și validității lor. Formele

alternative sunt deosebit de utile atunci când unei persoane trebuie să i se aplice același test în mai multe situații. *Efectele practice* (i.e. creșteri ale scorurilor care pot fi atribuite unei expuneri anterioare la itemii testului sau la itemi similari cu cei dintr-un test) intră în joc atunci când este dată o formă alternativă a unui test aplicat anterior, dar nu sunt la fel de mari ca atunci când aceeași versiune a testului este aplicată încă o dată.

• *Testele ancoră* constau din seturi comune de itemi administrate unor grupuri diferite de examinați în contextul a două sau mai multe teste și oferă o soluție diferită la problema comparării scorurilor obținute la test. Dacă avem răspunsurile mai multor grupuri normative la seturi comune de itemi, într-un cadru temporal dat, putem să extrapolăm și să comparăm scorurile unui test cu cele ale altuia atât pentru indivizi, cât și pentru grupuri. Această tehnică poate fi utilizată atunci când se urmărește facilitarea comparării nivelelor de performanță la diferite arii de abilități — cum ar fi înțelegerea unui text citit și exprimarea în scris — din două teste diferite, pe o scală uniformă. Cu toate acestea, în ultimii ani, este mult mai probabil ca acest tip de comparabilitate să fie atins fie prin normare simultană, fie prin tehnici ale teoriei răspunsului la itemi ce vor fi discutate pe scurt.

• *Grupurile fixe de referință* oferă o cale de a dobândi o anumită comparabilitate și continuitate a scorurilor testului în timp. Această metodă utilizează teste-ancoră încorporate în fiecare formă succesivă a testului pentru a oferi o legătură cu una dintre formele anterioare ale aceluiași test. În această manieră, o serie de teste este legată printr-un lanț de itemi comuni de scorurile grupului selectat ca referențial fix în scopul

menținerii continuității scalei scorurilor în timp. SAT este cel mai cunoscut exemplu de test care a utilizat grupuri fixe de referință. Până în aprilie 1995, toate scorurile SAT erau exprimate în termenii mediei și abaterii standard a 11000 de candidați la colegiu care dăduseră testul în 1941; pe acea scală, un scor de 500 corespundea mediei grupului de referință din 1941, un scor de 400 se găsea la 1 *SD* sub medie și așa mai departe. După aprilie 1995, scorurile SAT raportate reflectă o *recentrare* a scalei scorurilor pe candidați contemporani la colegiu, astfel încât un scor de 500 reprezintă nivelul mediu curent al performanțelor la test. Utilizarea unui grup de referință fix pentru SAT timp de câteva decenii a permis evaluarea creșterilor sau descreșterilor mărimii performanțelor candidaților la colegiu în epoci diferite, așa cum se poate vedea în Info 3.4. Pe lângă scorurile standard recentrate, scorurile rangului percentilic obținute la SAT de candidații la colegiu pot fi și sunt raportate utilizând ca grup de referință cel mai recent grup de absolvenți ai programelor de pregătire preuniversitară.

• *Normarea simultană* a două sau mai multe teste pe același eșantion de standardizare, deseori denumită co-normare, este încă o metodă utilizată pentru a obține scoruri ce pot fi comparate. Prin normarea testelor în același timp și pe același grup de persoane, se poate compara performanța indivizilor sau subgrupurilor la mai multe teste, utilizând același standard. Această posibilitate este deosebit de utilă atunci când se dorește contrastarea nivelelor de perfomanță relative pentru două sau mai multe funcții psihice, cum ar fi nivelele vocabularului expresiv și receptiv sau memoria de scurtă durată și memoria de lungă durată pentru același individ sau pentru un subgrup. Woodcock-Johnson III (WJ III) oferă un exemplu proeminent

Info 3.4

Schimbarea standardelor date de grupul de referință prin recentrarea SAT
Din anii '40 până în anii '90, scorurile candidaților la colegiu atât la probele verbale ale SAT, cât și la cele de matematică au prezentat scăderi semnificative. Astfel, după ce recentrarea a avut loc în anii '90, când a fost comparat cu grupul de referință din 1941, scorul Verbal recentrat de 500 s-a dovedit a fi echivalent cu un scor anterior de 420. Schimbarea reprezintă o scădere de aproape o unitate SD. Scorul la matematică de 500 s-a dovedit a fi echivalent cu un scor de 470 pentru grupul de referință din 1941.
Diferite argumente au fost aduse pentru aceste schimbări. Printre acestea, cel mai plauzibil se centrează pe diversitatea sporită a candidaților la colegiu din punctul de vedere al statutului socio-economic și etnic, precum și pe schimbările în calitatea curriculei învățământului gimnazial între începutul anilor '40 și începutul anilor '90.

de conormare a două baterii de teste. WJ III Tests of Cognitive Abilities (WJ III COG) este o baterie destinată să măsoare atât funcțiile cognitive generale, cât și pe cele specifice, în timp ce bateria WJ III Tests of Achievement (WJ III ACH) este destinată evaluării punctelor academice puternice și slabe ale unei persoane. Aceste două baterii au fost normate pe un eșantion extins de indivizi, de la preșcolari la adulți vârstnici, reprezentativ pentru populația Statelor Unite, oferind astfel un prilej pentru a compara intraindividual nivelele de performanță la câțiva indicatori ai funcționării cognitive și ai aptitudinilor academice.

Teoria răspunsului la item (TRI)

Diferite procedee sofisticate bazate pe modele matematice înlocuiesc din ce în ce mai mult tehnicile tradiționale de echivalare

care tocmai au fost descrise. Aceste procedee, care datează din anii '60 și sunt cunoscute drept *modele ale trăsăturii latente*, sunt grupate cel mai des sub eticheta *teoria răspunsului la item (TRI)*. Termenul *trăsătură latentă* reflectă faptul că aceste modele caută să estimeze nivele — neobservabile — ale diferitelor aptitudini, trăsături sau constructe psihologice care sunt subiacente comportamentului observabil al indivizilor, demonstrat de răspunsurile lor la itemii testului. Spre deosebire de tehnicile discutate anterior pentru echivalarea testelor și a scorurilor testelor, metodele TRI aplică modele matematice datelor *itemului* testului din eșantioane mari și diverse, de aici și denumirea *teoria răspunsului la item*. Aceste date sunt utilizate pentru a calibra itemii testului împreună cu unul sau mai mulți parametri și pentru a determina probabilitatea estimată a gradului aptitudinii, sau a nivelului trăsăturii, necesară pentru a răspunde la fiecare item într-un anumit fel. În mod esențial, aceste modele plasează atât persoanele, cât și itemii testului pe o scală comună (Embretson & Reise, 2000).

Elementele suplimentare și procedeele fundamentale din TRI vor fi discutate pe larg în Capitolul 6. Totuși, în contextul prezent merită subliniat faptul că, dacă modelele TRI îndeplinesc anumite condiții, ele pot produce estimări de parametri ai itemului care sunt *invariante* de-a lungul populațiilor. Aceasta înseamnă că aceste estimări nu sunt legate în mod necesar de performanța unui grup de referință specific. Datele răspunsurilor la item pot fi interpretate în termenii unei dimensiuni a aptitudinii sau a trăsăturii. Astfel, atunci când modelele TRI sunt aplicate unor arii de răspuns la item și datelor scorurilor la test din diferite eșantioane și sunt îndeplinite premizele modelelor, ele pot fi utilizate în două moduri reciproce: (a) pentru a estima probabilitatea ca persoanele cu anumite nivele specificate ale aptitudinii sau trăsăturii în

chestiune să răspundă corect la un item sau să răspundă la el într-un anumit mod și (b) pentru a estima nivelele trăsăturii necesare pentru a avea o probabilitate specificată de a răspunde la un item într-un anumit mod.

Testarea computerizată adaptativă

Unul dintre principalele avantaje ale metodologiei TRI constă în faptul că este ideală pentru utilizarea în testarea computerizată adaptativă (TCA). Atunci când indivizilor li se aplică teste computerizate adaptative (TCA), nivelele lor de aptitudine pot fi estimate pe baza răspunsurilor la itemii testului în timpul procesului de testare; aceste estimări sunt utilizate pentru a selecta itemii testului adecvați nivelului de aptitudine al persoanelor cărora li se aplică testul. Deși multe TCA au un număr stabilit de itemi, altele sunt construite astfel încât testarea se poate întrerupe ori de câte ori este îndeplinită o regulă specificată de oprire și nivelele trăsăturii sau aptitudinii au fost stabilite cu suficientă precizie. În orice caz, procedurile TCA reduc în mod semnificativ lungimea testului și timpul de testare; de asemenea, ele pot reduce frustrarea pe care o pot simți multe persoane care răspund la teste creion-hârtie atunci când sunt puse în fața broșurilor testului care conțin itemi ce se află considerabil peste nivelul lor de aptitudine sau sub acesta. Programele de testare pe scară largă, cum ar fi cele de la Educational Testing Service și U.S. Department of Defense, au încercat și au utilizat metodologia TCA timp de câțiva ani (vezi, e.g., Campbell & Knapp, 2001; Drasgow & Olson-Buchanan, 1999). Deși această metodologie are avantaje clare în raport cu testele creion-hârtie cu dimensiune fixă — și aria lor de aplicare tinde să se extindă — ele prezintă unele probleme noi legate de securitatea testului, de costurile testului și

de incapacitatea celor examinați de a-și revedea și a-și modifica răspunsurile (Wainer, 2000).

Revizuirile unui test

Denumirea testului poate reprezenta cu ușurință o posibilă sursă de confuzie pentru potențialii utilizatori ai testului, insuficient informați. Rareori se justifică să te orientezi după titlurile testelor. Unele discrepanțe între titlurile testelor și ceea ce testele evaluează de fapt sunt destul de evidente: inventarele de personalitate, de fapt, nu inventariază personalitatea. Altele nu sunt atât de evidente: testele de aptitudini pot sau nu să testeze aptitudinile și așa mai departe. Chiar dacă testele au același nume, ele pot să nu fie echivalente; numeroase teste sunt revizuite din când în când și păstrează același titlu, cu excepția adăugării unui număr sau a unei litere care identifică o anumită versiune (ex.: WISC, WISC-R, WISC-III, WISC-IV). În general, cu cât testele au succes mai mare și sunt mai larg utilizate, este probabil să fie revizuite mai des decât alte teste. Obiectivele și măsura revizuirii testului se pot întinde de la schimbări minore ale cuvintelor itemilor, la reorganizări majore ale conținutului, scorării sau ale procedurilor de administrare.

Un test care a fost modificat într-un anumit mod nu poate fi considerat comparabil cu o versiune anterioară, în afara cazului în care similaritatea lor este stabilită empiric. Atunci când revizuirea unui test este minoră și nu se așteaptă o influențare a scorurilor, măsura în care pot fi comparate versiunea veche și cea revizuită poate fi stabilită relativ ușor. De obicei, aceasta se realizează aplicând ambele versiuni aceluiași grup de persoane. Dacă cele două versiuni corelează puternic și au medii, abateri

standard și distribuții ale scorurilor similare, în scopuri practice se presupune că versiunile sunt interșanjabile. Un caz tipic al acestui procedeu este transferarea testelor creion-hârtie, cum ar fi MMPI-2 într-un format de administrare computerizată fără nicio schimbare a conținutului sau scorării testului.

Pe de altă parte, revizuirile majore ale testelor cu raportare la normă necesită restandardizarea testului cu un nou eșantion normativ. Astfel, atunci când schimbările sunt semnificativ suficiente pentru a justifica diferențele în scala sau scorarea testului, avem de-a face, de fapt, cu un test nou și diferit, deși testul poate păstra o anumită asemănare cu versiunile anterioare și ar putea avea același titlu. Un exemplu ilustrativ este Scala de inteligență Stanford-Binet (S-B) care a fost publicată pentru prima dată în 1916. Cea de-a patra și a cincea ediție a S-B, publicate în 1986 și, respectiv, 2003, sunt diferite de versiunile sale anterioare în aproape toate aspectele și, în timp, au devenit mai apropiate ca formă și conținut de scalele Wechsler decât de S-B original.

Schimbări longitudinale ale normelor testului

Atunci când un test este revizuit și standardizat pe un nou eșantion după o perioadă de câțiva ani, chiar dacă revizuirile conținutului său sunt minore, normele de scorare tind să se deplaseze într-o direcție sau alta datorită schimbărilor apărute în populație după diferite perioade de timp. O astfel de schimbare, discutată în Info 3.4, e reprezentată de scăderea scorurilor medii SAT între grupul fix de referință testat în 1941 la candidații la colegiu din anii '90. O tendință uimitoare în direcția opusă, cunoscută ca *efectul Flynn*, a fost bine documentată în revizuirile succesive ale testelor importante de inteligență (ca scalele S-B și Wechsler) care, invariabil, presupun atât administrarea versiunilor

noi, cât și a celor vechi unui segment al noului eșantion de standardizare, în scopul comparării acestora. Datele obținute din revizuirea diferitelor teste de inteligență în Statele Unite și în alte țări — analizate extensiv de J. R. Flynn (1984, 1987) — prezintă o tendință ascendentă pronunțată, de lungă durată, a nivelului de performanță necesar pentru a obține un scor IQ dat. Este de presupus că efectul Flynn reflectă achizițiile populației de-a lungul timpului în aria tipului de performanță cognitivă pe care testele de inteligență o reprezintă. Diferiți factori — cum ar fi nutriția îmbunătățită, îngrijirile prenatale mai bune și o mai mare complexitate a mediului — au fost propuși drept motive ale acestor rezultate. Însă, măsura în care efectul Flynn se generalizează la toate populațiile din lume — ca și cauzele sale posibile — rămâne un subiect al unei controverse importante (Neisser, 1998).

Un exemplu relevant de schimbări longitudinale în performanța la un test de personalitate a fost întâlnit în renormarea MMPI (publicat inițial în anii '40) care a avut loc în anii '80 ca parte a construirii MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Modificările conținutului inventarului au fost relativ minore. Cu toate acestea, trecerea de la normele inițiale MMPI la cele ale MMPI-2 au dus la modificări ale nivelurilor scorurilor la diferite scale, considerate semnificative din punct de vedere clinic datorită schimbărilor substanțiale ale modului în care au răspuns la item oamenii din cele două perioade diferite de timp. Tipul de schimbări tocmai descris este obișnuit atunci când eșantioanele normative sunt aduse la zi. Ele subliniază necesitatea unor astfel de aduceri la zi pentru a menține actualitatea normelor din interiorul grupului. Astfel de schimbări explică, de asemenea, motivul pentru care documentația ce

Aplicație practică

Cum poate influența „efectul Flynn" schimbările aparente ale scorurilor obținute la un test de inteligență: un exemplu utilizând Scala de inteligență Wechsler pentru copii (WISC)

John a obținut un scor IQ Wechsler de 117 la WISC-R la vârsta de 9 ani și un IQ Wechsler de 107 la WISC-III la vârsta de 13 ani. Acest declin semnificativ al performanței sale nu poate fi luat după aparențe.

Un factor care ar putea explica parțial acest declin este efectul Flynn. Acest efect se referă la cel mai înalt nivel de performanță care este văzut în mod tipic în grupurile normative ale versiunilor mai noi ale testelor de inteligență generală comparativ cu pandantele lor mai vechi — de exemplu, WISC-III, publicat în 1991, comparat cu WISC-R publicat în 1974. Conform manualului WISC-III, publicat în 1991, unui eșantion alcătuit din 206 copii cu vârsta cuprinsă între 6 și 16 ani i s-a aplicat WISC-R și WISC-III, contrabalansat, la intervale între 12 și 70 de zile între cele două testări. IQ mediu pe Scala completă a fost 108,2 la WISC-R și 102,9 la WISC-III (Wechsler, 1991). Deși corelația dintre cele două versiuni ale WISC pentru acest eșantion a fost puternică (0,89), scorurile copiilor la WISC-III au fost în medie mai scăzute — cu ceva mai mult de 5 puncte — decât cele la WISC-R. Diferența scorurilor IQ medii indică faptul că în raport cu normele celor două teste, în general, orice scor IQ dat reprezintă un nivel de performanță mai înalt la testul mai nou decât la cel mai vechi.

Bineînțeles, mulți factori, pe lângă efectul Flynn, ar fi putut contribui la diferența obținută între cele două IQ-uri ale lui John. Pe un loc principal se află eroarea de măsură a fiecărui scor (vezi Capitolul 4) și posibilitatea ca John să fi traversat un declin real în funcționarea intelectuală generală datorită unei anumite boli sau a unei situații de viață.

însoțește orice tip de test cu raportare la normă ar trebui să ofere o descriere completă și precisă a componenței eșantioanelor utilizate pentru a-l standardiza și norma, inclusiv momentul în care

a fost format eșantionul. Documentația testului trebuie să cuprindă răspunsuri la toate întrebările enumerate în Info 3.2, incluzând o descriere clară a pașilor urmați în construirea testului și a procedeelor de administrare și scorare din timpul standardizării sale. Utilizatorii testelor, la rândul lor, trebuie să cunoască această informație și să țină cont de ea atunci când selectează teste și interpretează scoruri ale testelor.

Interpretarea testelor în raport cu un criteriu

În domeniul evaluării educaționale și ocupaționale, testele sunt utilizate deseori pentru a ne ajuta să stabilim dacă o persoană a atins un anumit nivel de competență într-un domeniu de cunoaștere sau a dobândit o anumită abilitate în realizarea unei sarcini. În astfel de cazuri, cadrul de referință pentru interpretarea testelor trebuie schimbat. În loc să comparăm performanța persoanei cu cea a celorlalte, performanța unui individ sau a unui grup este comparată cu un criteriu sau un standard prestabilit. Atunci când este utilizat în acest context, *criteriul* se poate referi fie la cunoașterea unui domeniu de conținut specific, fie la competența într-un anumit tip de demers. Standardele cu care sunt evaluate testele cu raportare la criteriu sunt definite în mod tipic în termeni de nivele specificate de cunoștințe sau expertiză necesare pentru a absolvi un curs, pentru a obține un grad sau pentru a dobândi o licență profesională; ele pot presupune demonstrarea unei competențe suficiente pentru a efectua o activitate sau pentru a realiza un produs. Deseori, dar nu întotdeauna, aplicarea testelor cu raportare la criteriu presupune utilizarea scorurilor limită sau a intervalelor de scoruri care separă competența de incompetență sau delimitează nivele diferite de performanță. În astfel de cazuri,

validitatea inferenţelor realizate pe baza scorurilor trebuie să fie stabilită prin legături empirice între scorurile obţinute la teste şi performanţa atinsă la criteriu.

Tipuri de interpretare a testului în raport cu un criteriu

Termenul *testare în raport cu criteriul* popularizat de Glaser (1963) este uneori utilizat ca sinonim cu testarea *în raport cu domeniul, în raport cu conţinutul, cu referinţă obiectivă* sau testarea *competenţei*. Această stare de lucruri mai degrabă confuză se datorează, în parte, faptului că interpretarea testului în raport cu un criteriu utilizează cel puţin două seturi de standarde subiacente: (a) cele bazate pe *gradul de cunoaştere* a conţinutului unui domeniu demonstrat de testele obiective standardizate şi (b) cele bazate pe *nivelul de competenţă* într-o arie de abilităţi prezent în calitatea performanţei sau a produsului care a rezultat în urma exercitării acelei abilităţi. Ocazional, termenul *testare în raport cu criteriul* este utilizat şi pentru a ne referi la interpretările bazate pe relaţii prestabilite între scorurile obţinute la un test şi nivelele aşteptate de performanţă în raport cu un criteriu, cum ar fi o activitate viitoare sau chiar şi un alt test. Cu această utilizare specială, „criteriul" este un rezultat specific şi ar putea să fie sau să nu fie legat de sarcinile al căror eşantion este testul. Acesta din urmă este în contrast puternic cu testele cu raportare la conţinut sau cu testele de performanţă, în care sarcinile testului sunt în mod esenţial eşantioane de comportament în directă relaţie cu criteriul. Info 3.5 enumeră câteva puncte majore în care testarea în raport cu norma diferă de testarea în raport cu criteriul, ca şi unele diferenţe între tipurile de testare în raport cu criteriul.

Info 3.5

Interpretarea testului în raport cu criteriul versus norma

- Testele *cu raportare la normă* urmăresc să localizeze performanţa unuia sau mai multor indivizi la constructul evaluat pe un continuum creat de performanţa unui grup de referinţă.
- Testele *cu raportare la criteriu* urmăresc să evalueze performanţa indivizilor în relaţie cu standarde referitoare la constructul în sine.
- În timp ce în interpretarea testelor cu raportare la normă cadrul de referinţă este dat întotdeauna de oameni, în interpretarea testelor cu raportare la criteriu cadrul de referinţă poate fi
  - cunoaşterea unui domeniu de conţinut aşa cum se demonstrează în testele standardizate, obiective; sau
  - nivelul de competenţă prezent în calitatea unei performanţe sau a unui produs
- Termenul *testare în raport cu criteriul* este uneori folosit şi pentru a descrie interpretările testelor care utilizează drept cadru de referinţă relaţia dintre scoruri şi nivele aşteptate de performanţă sau poziţii în raport cu un criteriu.

Pe lângă distincţiile anterioare, variază şi maniera particulară în care sunt utilizate testele în raport cu un criteriu. Uneori criteriile sunt strict cantitative, ca atunci când este nevoie de anumite procente (80% sau 90%) de răspunsuri corecte pentru a stabili o măiestrie adecvată. Alteori, criteriile au o natură mai mult calitativă şi subiectivă. Mai mult, uneori performanţa la aceste teste este evaluată pe o bază de tipul totul-sau-nimic ca răspuns la întrebarea dacă a fost dobândit un anumit nivel de competenţă, iar alteori sunt permise nivele intermediare de competenţă.

În ciuda diferenţelor în privinţa accentului şi denumirii testelor cu raportare la criteriu, aceste instrumente au anumite caracteristici comune. În mod tipic, testele cu referire la criteriu (a) stabilesc

măsura în care persoanele testate sunt competente în anumite domenii de abilități sau de cunoștințe și (b) sunt scorate astfel încât performanța unei persoane nu influențează poziția celorlalte. În timp ce testele cu raportare la normă caută să ierarhizeze sau să pună un individ în relație cu ceilalți din perspectiva constructului pe care îl evaluează, testele cu raportare la criteriu caută să evalueze performanța indivizilor în relație cu constructul respectiv.

În contextul prezent, doar acele aspecte ale interpretării testelor cu raportare la criteriu necesare pentru o înțelegere bazală a premiselor sale și a terminologiei vor fi discutate. Câteva dintre aceste concepte sunt reluate mai detaliat în Capitolul 5 în legătură cu tema validității, de care sunt legate inextricabil. Info 3.6 enumeră câteva surse de informație din vasta literatură disponibilă despre diferite tipuri de testare în raport cu criteriul.

Testarea cunoștințelor din diferite domenii de conținut

Pentru a utiliza cunoașterea domeniilor de conținut drept cadru de referință al interpretării scorurilor obținute la test, trebuie să existe un câmp sau un subiect definit cu atenție și demarcat în mod clar din care să se extragă eșantioanele (i.e. itemi ai testului sau sarcini) pentru a estima cunoștințele celui căruia îi este aplicat testul. Mai mult, obiectivele trebuie evaluate atât în termenii cunoașterii unui domeniu de conținut, cât și în cei ai aplicării acestor cunoștințe și ai standardelor care vor fi utilizate în evaluarea acestor obiective, care ar trebui să derive dintr-un consens al persoanelor experte în domeniu. O astfel de situație este întâlnită în principal în cadrul educațional sau de formare, în care subiectele și disciplinele tind să fie împărțite în lecții, cursuri, programe de studiu și alte unități curriculare predate elevilor și

Info 3.6

Selecție de lucrări despre testarea în raport cu un criteriu

Literatura referitoare la testarea în raport cu un criteriu, datând din anii '60, este bogată. Pentru a urmări pe larg această temă, cititorii pot consulta una sau mai multe dintre următoarele surse:

- Cartea lui Cizek (2001) despre *Setting Performance Standards* se concentrează atât pe aspectele teoretice, cât și pe cele practice ale stabilirii standardului și pe ramificațiile sale numeroase.
- Articolul clasic al lui Popham și Husek (1969) despre implicațiile măsurătorilor în raport cu un criteriu;
- Capitolul scris de Hambletin și Rogers (1991) despre „Advances in Criterion-Referenced Measurement" oferă o introducere utilă în domeniu și o descriere a progreselor tehnice între anii '60 și '80.
- Volumul editat de Wigdor și Green (1991) despre diferite aspecte ale evaluării performanței la locul de muncă și
- Clarificarea concisă, dar lămuritoare, a lui Linn (1994) privind unele confuzii ale diferitelor interpretări ale testării în raport cu un criteriu.

Referințele complete ale acestor lucrări sunt enumerate în secțiunea de bibliografie de la sfârșitul acestei cărți.

din care pot fi eșantionate arii de conținut și rezultate ale învățării. Aceste teste sunt descrise de obicei ca măsuri ale „achiziției de cunoștințe". Ele tind să aibă itemi — cum ar fi întrebări cu răspunsuri multiple — care cer persoanelor testate să selecteze un răspuns sau să îndeplinească o sarcină cu un grad înalt de structurare (cum este scrierea unui scurt paragraf pe o anumită temă sau rezolvarea unei probleme matematice).

Atunci când domeniile de cunoaștere reprezintă cadrul de referință al interpretării testului, întrebarea la care trebuie să se

răspundă este „Cât de mult din domeniul specificat stăpânește cel căruia i-a fost aplicat testul?" și scorurile sunt prezentate cel mai des sub forma procentului de răspunsuri corecte. Acest tip de interpretare a testului în raport cu criteriul este deseori descrisă drept *testare în raport cu conținutul* sau *cu domeniul*. De fapt, unii consideră acești doi termeni ca fiind sinonimi cu testarea în raport cu un criteriu. Planificarea acestor teste necesită dezvoltarea unui *tabel de specificații* cu celule care specifică numărul de itemi sau sarcini ce vor fi incluse în test pentru fiecare dintre obiectivele de învățare și ariile de conținut pe care testul este destinat să le evalueze. Proporția de itemi ai testului alocată fiecărei celule reflectă greutatea sau importanța atribuită fiecărui obiectiv și fiecărei arii. Info 3.7 prezintă exemple de diferite obiective și itemi din două arii de conținut tipice pentru testele cu referință la domeniu. Exemple de tabele de specificații pentru testele cu raportare la criteriu și indicații pentru modul în care acestea sunt pregătite pot fi găsite în Gronlund (2003) și Linn și Gronlund (1995, pp. 119–125).

Evaluarea performanțelor

De asemenea, pentru luarea deciziilor în domeniul ocupațional și în cel educațional, există deseori nevoia de a stabili sau a certifica competențe în realizarea unor sarcini care să fie mai realiste, mai complexe, care consumă mai mult timp sau sunt mai dificil de evaluat decât cele tipice pentru testarea în raport cu conținutul sau domeniul. Acest tip de evaluare necesită evaluarea performanței prin eșantioane de activitate de muncă, produse ale muncii sau alte manifestări comportamentale ale competenței și abilității în situații care simulează cadrele din viața reală.

Info 3.7

Exemple de obiective și itemi ai unui test cu raportare la domeniu

I. *Domeniu:* Aritmetică
   A. *Aria de conținut ce va fi evaluată:* Înmulțirea fracțiilor
   B. *Obiectivele ce vor fi evaluate:*
      1. Cunoașterea pașilor implicați în înmulțirea fracțiilor
      2. Înțelegerea principiilor de bază implicate în înmulțirea fracțiilor
      3. Aplicarea principiilor în rezolvarea problemelor de înmulțire a fracțiilor
   C. *Eșantion de itemi ai testului pentru fiecare obiectiv:*
      Item 1. Enumerați pașii implicați în înmulțirea fracțiilor
      Item 2. Desenați o diagramă pentru a indica 1/4 din 1/2 dintr-un grafic circular.
      Item 3. Cât face 3/4 x 1/2?

II. *Domeniu:* Vocabular
   A. *Aria de conținut ce urmează a fi evaluată:* Cunoașterea cuvintelor
   B. *Obiectivele ce vor fi evaluate:*
      1. Definiția cuvântului
      2. Înțelegerea sensului cuvântului
      3. Aplicarea cunoașterii cuvântului în exprimarea scrisă
   C. *Eșantion de itemi ai testului pentru fiecare obiectiv:*
      Item 1. Ce înseamnă „marinar" _____
      Item 2. Care cuvânt este cel mai apropiat de sensul cuvântului „marinar"?
         a. Marinat
         b. Marimba
         c. Navigator
         d. Pirat
         e. Rătăcitor
      Item 3. Formați o propoziție utilizând „marinar" într-un context cu sens.

Atunci când scopul unei evaluări este de a stabili nivele de competență în tipul de contexte în care abilitățile sunt aplicate în viața reală, criteriul din interpretarea testului „în raport cu un criteriu" este fie calitatea performanței, fie a produsului care rezultă din punerea în practică a acelei abilități. În acest cadru de referință, întrebările la care trebuie să se răspundă sunt: „Persoana testată deține abilitatea în chestiune?" sau „Cât de competentă este această persoană testată, sau acest grup, în continuumul de competență relevant pentru această sarcină particulară?"

Evaluarea și scorarea în determinarea performanței

În lumina întrebărilor la care trebuie să răspundă evaluarea competenței, determinarea performanței impune un set diferit de procedee decât cele utilizate atunci când se realizează o testare a cunoașterii domeniilor de conținut. În general, evaluarea performanței tinde să se bazeze mult mai puternic pe judecată subiectivă. O excepție de la această regulă apare atunci când criteriile pot fi cuantificate în termenii rapidității realizării, numărului de erori, de unități produse sau alte standarde obiective. Un exemplu clasic și simplu de tip obiectiv de evaluare a performanței este testul de dactilografiere aplicat candidaților la servicii de funcționari care necesită în mare măsură dactilografierea. Totuși, chiar și în acest tip de test, criteriul numeric concret utilizat pentru a separa scorurile de performanță acceptabilă — de exemplu 65 de cuvinte pe minut și nu mai mult de 5 erori — este probabil să fie stabilit arbitrar. Multe alte tipuri de evaluare a performanței presupun (a) identificarea și descrierea criteriilor calitative de evaluare a performanței sau produsului și (b) construirea unei metode de aplicare a criteriilor. Metodele obișnuite de evaluare a criteriilor calitative presupun *scale de ierarhizare* sau *rubrici de scorare* (i.e. ghiduri de scorare) care descriu și ilustrează regulile și principiile ce urmează a fi aplicate în scorarea calității unei performanțe sau a unui produs. Un exemplu binecunoscut al acestui tip de procedură este scorarea performanțelor atletice de către evaluatori experți numiți în cadrul unor evenimente cum ar fi patinaj artistic sau competiții de sărituri în apă.

Testarea îndemânării. Procedurile care evaluează performanța la test pe baza faptului că un subiect demonstrează sau nu un nivel prestabilit de îndemânare sunt cunoscute drept *teste de îndemânare.* Multe dintre aceste teste necesită scorări de tipul totul-sau-nimic, cum ar fi *corect* sau *greșit*, bazate pe un anumit nivel al criteriului care separă îndemânarea de neîndemânare. Un exemplu tipic, familiar pentru cei mai mulți cititori, este oferit de testele de conducere auto pe care multe state le cer pentru a emite permisul de conducere. În aceste teste, ceea ce contează este dacă indivizii pot demonstra că ei cunosc regulile de conducere și sunt capabili să manevreze un automobil în diferite situații de trafic. Mai mult, se așteaptă ca majoritatea persoanelor care dau un test de conducere să fie capabile să îl treacă, deși sunt necesare mai multe încercări, și nu e nevoie să se facă nicio altă distincție între cei care dau testul cu privire la nivelul lor de performanță, alta decât că au trecut sau nu testul.

Există unele situații și abilități — cum ar fi aterizarea unui avion pe un portavion sau realizarea unei operații pe creier — în care erorile în determinarea exactă a îndemânării desăvârșite sunt inacceptabile. Pe de altă parte, în cercurile educaționale, noțiunea de testare a îndemânării poate fi interpretată în moduri diferite. Unii educatori și alte părți interesate au susținut că toți studenții capabili trebuie să dobândească stăpânirea desăvârșită a obiectivelor instrucționale prescrise la un anumit nivel înainte

de a absolvi și a trece la următorul nivel, indiferent cât de mult timp este necesar. Cu toate acestea, cei mai mulți pot să admită că, atunci când testarea este destinată evaluării stăpânirii abilităților de bază predate în școală — cum ar fi cititul, scrisul și aritmetica — există în mod clar un loc pentru diferite nivele de achiziție între stăpânirea materiei și absența acesteia. În astfel de cazuri, problema este aceea de a stabili obiective adecvate, pe care elevii sunt așteptați să le atingă, în cadrul unui continuum de performanță, pentru a promova sau pentru a absolvi.

Predicția performanței

Uneori, termenul *interpretarea testului în raport cu un criteriu* este utilizat pentru a descrie aplicarea datelor empirice cu privire la legătura dintre scorurile obținute la test și nivelele de performanță la un criteriu cum ar fi performanța la locul de muncă sau succesul într-un program de studii. În acest context, termenul *criteriu* este utilizat într-un sens diferit și într-un acord mai mare cu practica psihometrică tradițională decât în exemplele precedente. Aici, criteriul este un rezultat care trebuie estimat și prezis prin intermediul unui test. Acest tip de informație constituie baza pentru stabilirea validității predictive a testelor ce va fi discutată în detaliu în Capitolul 5. Cu toate acestea, este menționată aici deoarece, atunci când este utilizată o relație între scorurile testului și criterii pentru selectarea sau plasarea indivizilor în cadre educaționale, de formare sau de muncă, acea relație trebuie să fie construită și ca un cadru de referință al interpretării scorurilor. În acest cadru, întrebările la care se răspunde cu ajutorul scorurilor testului sunt „Ce nivel al performanței la criteriu se poate aștepta de la o persoană care obține acest scor?" sau „Performanța la test este suficientă pentru a asigura

nivelul dorit de performanță la criteriu într-o sarcină dată?". Informațiile despre relația dintre scorurile testului și criterii pot fi prezentate în diferite moduri, incluzând coeficienți de corelație și ecuații de regresie care sunt descrise mai pe larg în contextul validității în Capitolul 5. Cu toate acestea, două procedee deosebit de relevante pentru tipul de interpretare a testului în raport cu criteriul discutată în paragraful precedent — și anume tabele ale expectațiilor și grafice de expectație — vor servi la clarificarea acestei abordări.

Tabelele de expectanță prezintă distribuția scorurilor testului pentru unul sau mai multe grupuri de indivizi, tabelată încrucișat cu performanța la criteriu. Presupunând că există un grad de corelație substanțial între scorurile la test și măsurile criteriului, această informație poate fi utilizată pentru a estima poziția probabilă la criteriu a indivizilor care au obținut scoruri la diferite niveluri. De exemplu, utilizând informația din clasele anterioare, un profesor ar putea nota într-un tabel scorurile la testul parțial ca predictor și nota finală la curs drept criteriu, așa cum este prezentat în Tabelul 3.3. Pe baza scorurilor obținute la testul parțial, tabelul rezultat poate informa viitorii elevi cu privire la nota finală probabilă.

Graficele de expectanță sunt utilizate atunci când performanța la criteriu într-o sarcină de muncă, un program de formare sau de studiu poate fi clasificată drept reușită sau nereușită. Aceste grafice reprezintă distribuția scorurilor unui grup de indivizi alături de procentul de oameni din fiecare interval de scoruri care au reușit (sau au ratat) în raport cu criteriul. Atunci când tendința este astfel încât procentul de indivizi care au succes este mult mai mare printre cei cu scoruri înalte decât printre cei cu scoruri scăzute, graficele de acest tip pot fi extrem de utile în luarea deciziilor de selecție.

Tabelul 3.3 Tabel de expectanță: relația dintre scorurile la testul parțial și notele finale

Scorurile la testul parțial	Număr de cazuri	Procentul care a primit fiecare notă finală			
		D sau F	C	B	A
90 și peste	9		11	22	67
De la 80 la 89	12	8	25	50	17
De la 70 la 79	13	23	46	31	
De la 60 la 69	3	33	67		
59 și sub	3	67	33		

Relația dintre cadrele de referință

Nu se poate face distincția dintre testele cu raportare la normă și cele cu raportare la criteriu pur și simplu, privind exemplare din fiecare tip. De fapt, într-o mare măsură, distincția dintre cele două cadre de referință ale interpretării unui test — ca și dintre tipurile diferite din fiecare cadru — ține de modul în care este pus accentul. Chiar dacă scorurile pot fi exprimate în diferite forme, în mod fundamental, testarea se bazează pe un cadru normativ.

Standardele utilizate în interpretările scorurilor unui test în raport cu un criteriu trebuie să se bazeze pe așteptări realiste sau fezabile pentru populația sau persoanele pentru care este construit testul. În funcție de scopurile în care este utilizat un test, pot exista foarte puțini sau foarte mulți oameni capabili să satisfacă un criteriu, caz în care testul se poate dovedi a fi nepractic. Cu alte cuvinte, criteriile se bazează atât pe scopul testării, cât și, într-o anumită măsură, pe ceea ce oamenii pot realiza într-o situație dată.

Utilizarea scorurilor — prag sau a altor standarde de performanță în interpretarea testului în raport cu criteriul nu înseamnă că diferențele performanțelor la test ale indivizilor vor fi eliminate sau vor trece neobservate, și nici nu împiedică compararea scorurilor între indivizi. De exemplu, dacă doi candidați la un post de secretar satisfac criteriul de 65 de cuvinte pe minut la un test de dactilografiere, toate celelalte aspecte rămânând neschimbate, este mai probabil ca cel care poate dactilografia 90 de cuvinte pe minut să fie ales în detrimentul celui care poate dactilografia numai 65.

În mod similar, utilizarea normelor nu împiedică examinarea performanței la test din punctul de vedere al criteriilor de conținut sau comportamentale. În scorarea testelor aplicate în clasă, de exemplu, profesorii care au eșantionat cu atenție materialul testului pot alege să atribuie note pe baza normelor (pe o gradare curbilinie) sau pe baza criteriilor de performanță legate de procentul de itemi la care s-a răspuns corect. Unele teste standardizate utilizate în cadre educaționale — cum ar fi Iowa Tests of Basic Skills (ITBS), Stanford Diagnostic Mathematics Test (SDMT) și Stanford Diagnostic Reading Test (SDRT) — încearcă, de asemenea, să ofere atât interpretări în raport cu un criteriu, cât și în raport cu o normă. Cu toate acestea, unele recenzii ale acestor teste sugerează că această încercare are succes numai pentru un tip de interpretare sau altul, dar nu pentru ambele, deoarece este necesar un accent oarecum diferit în construirea testelor cu raportare la criteriu față de cele cu raportare la normă. Pentru mai multe informații despre această problemă, consultați recenziile la ITBS, SDMT și SDRT din a treia ediție a *Mental Measurements Yearbook*, editată de Impara și Plake (1998).

Atunci cum diferă interpretarea unui test raportat la normă de interpretarea testului raportat la un criteriu? Diferența fundamentală dintre cele două constă în obiectivele lor principale:

- În testarea raportată la normă, obiectivul principal este de a face distincții între indivizi în termenii abilității sau trăsăturii evaluate de un test;
- În testarea raportată la criteriu, obiectivul principal este de a evalua gradul de competență al unei persoane, de stăpânire a unei abilități sau a unui domeniu de cunoștințe în termenii unui standard de performanță prestabilit.

Așa cum am văzut, aceste două obiective nu se exclud reciproc întotdeauna sau nu în mod necesar și, în anumite situații, același instrument poate fi utilizat pentru ambele obiective. Scopul specific al utilizatorului testului determină principalul obiectiv dintre cele două. Acest scop, la rândul său, va ajuta utilizatorul testului să determine care abordare a construirii testului este cea mai potrivită.

Cu privire la diferitele tipuri de interpretare a testului în raport cu criteriul, ar trebui să fie destul de clar faptul că, într-o anumită măsură, distincția dintre evaluarea bazată pe domeniu și evaluarea bazată pe performanță este, de asemenea, arbitrară. Cunoașterea domeniilor de conținut trebuie să fie demonstrată prin comportamentul observabil sau prin performanța în care una sau mai multe abilități joacă un rol. Din același motiv, orice tip de performanță necesită un anumit tip de cunoștințe. Mai mult, dacă subiectele elementare și domeniile de conținut destul de bine structurate pot fi trasate în detaliu, când este vorba despre arii mai avansate și mai puțin structurate — cum ar fi cunoștințe care

intersectează mai multe discipline — acest tip de divizare devine dificil sau imposibil de realizat. În mod similar, pragurile de îndemânare pot fi prestabilite cu ușurință în privința abilităților de bază, dar în domenii care presupun abilități de nivel mai înalt, nu se pot stabili astfel de standarde deoarece achizițiile au o extindere mult mai largă.

Interpretarea testelor în raport cu un criteriu în evaluarea clinică

Criterii cum ar fi stăpânirea unei abilități sau a cunoștințelor dintr-un domeniu dat nu pot fi aplicate în mod clar în cazul instrumentelor destinate evaluării personalității. Prin urmare, termenul *interpretare în raport cu un criteriu* nu este utilizat de obicei pentru teste de acest tip. Cu toate acestea, unele teste relevante atât pentru funcționarea emoțională, cât și pentru cea cognitivă sunt utilizate de clinicieni pentru a stabili dacă anumite criterii de diagnostic au fost sau nu îndeplinite. Aceste teste utilizează scoruri limită — stabilite pe baza datelor normative — pentru a tria persoanele în funcție de prezența anumitor tulburări pe baza unor legături stabile cu criteriile clinice. În această aplicație particulară, interpretarea în raport cu criteriul e utilizată asemenea situației în care relația dintre scorurile obținute la test și criteriu este utilizată pentru a selecționa sau pentru a plasa indivizi în cadru educațional sau ocupațional. Ambele practici presupun estimarea sau predicția anumitor rezultate. De asemenea, acest tip de interpretare a testelor clinice depinde de validitatea dovezilor relevante pentru criteriile de diagnostic (vezi capitolul 5). Exemple de teste care sunt utilizate în această manieră includ instrumente cum ar fi Inventarul de Depresie Beck, care poate

ajuta la evaluarea intensității tulburărilor depresive și Mini-Mental State Examination, care ajută la identificarea deteriorărilor cognitive. Aceste teste și alte instrumente clinice — cum ar fi listele de comportamente și scalele de evaluare — ce presupun utilizarea scorurilor limită și a intervalelor de scoruri pentru a evalua comportamentul simptomatic al tulburărilor mentale pot fi considerate și teste cu raportare la criteriu.

Teoria răspunsului la item ca bază pentru combinarea cadrelor de referință

Deoarece scopul metodelor TRI este de a estima poziția subiectului testat în privința unei trăsături latente sau dimensiunea unei aptitudini, ele sunt potrivite pentru construirea testelor ale căror scoruri pot fi interpretate atât pe baze normative, cât și pe baza criteriului. Deși datele utilizate în modelarea TRI sunt obținute din performanța eșantioanelor de referință, ele pot fi combinate cu alte tipuri de analiză de item pentru a crea scale care oferă informații de natură comparativă (i.e. cu raportare la normă) și substanțială (i.e. cu raportare la criteriu). Un exemplu recent al modului în care se poate realiza aceasta poate fi găsit în lucrarea lui Primi (2002), care integrează teoria cognitivă și TRI în construirea unui instrument pentru măsurarea tipului de abilitate necesară pentru a rezolva probleme de geometrie matriceală. Un alt exemplu este modelul ierarhic TRI propus de Janssen, Tuerlinckx, Meulders și De Boeck (2000), care este aplicat unui test care măsoară atingerea obiectivelor în înțelegerea lecturii la nivelul școlii elementare.

Considerații sociale în testarea în raport cu un criteriu și cu o normă

Presiunile curente pentru utilizarea testării în raport cu un criteriu pentru certificarea competențelor și luarea deciziilor în cadrele educaționale și ocupaționale sunt cauzate de nemulțumirea legată de perceperea punctelor slabe ale testării cu raportare la normă. Unele dintre aceste nemulțumiri au apărut deoarece utilizarea raportării la normă în educație este o cauză majoră a scăderii standardelor, dat fiind că, indiferent cât de slabă ar fi performanța unei populații de elevi ca întreg, relativ la propriile lor norme, cel puțin jumătate dintre elevi vor fi întotdeauna peste medie. O altă sursă de nemulțumire față de testarea raportată la normă derivă din faptul că, atunci când este utilizată ca bază pentru luarea deciziilor în domeniile educațional și ocupațional, deseori dezavantajează membrii grupurilor minoritare neprivilegiate în comparație cu cei care au putut avea mai multe oportunități educaționale. Cu toate acestea, ironia face ca în timpul ultimelor câteva decenii utilizarea testării cu raportare la criteriu pentru certificarea competențelor a devenit subiectul unei dezbateri mai aprinse în arenele profesionale și politice decât testarea cu raportare la normă. Fără îndoială, mare parte din dezbaterea despre ambele tipuri de testare depinde de înțelegerea greșită a rolului adecvat al testelor ca instrumente — și nu ca arbitri — ca și a principiilor care par să alterneze între două extreme, ale supraaccentuării testării standardizate sau a opoziției față de aceasta care a ajuns să fie cunoscută drept „decizie dificilă" (vezi Jaeger, 1989; Mehrens, 1992; U.S. Department of Education, Office for Civil Rights, 2000; Wigdor & Green, 1991).

Testați-vă singuri

1. **Dacă nu este însoțit de o informație suplimentară, un scor brut înalt este**
   a. Lipsit de sens
   b. Întotdeauna mai bun decât un scor scăzut

2. **_____ constituie cel mai larg utilizat cadru de referință pentru interpretarea scorurilor.**
   a. Domenii de conținut
   b. Eșantioane ale activității
   c. Criterii
   d. Norme

3. **Dintre toate normele următoare de dezvoltare, care sunt aplicabile în mod universal?**
   a. scalele la nivel ordinal cu bază teoretică
   b. normele de vârstă mentală
   c. secvențe naturale
   d. norme bazate pe clase

4. **Cu privire la eșantioanele utilizate pentru a stabili norme în interiorul grupurilor, cea mai importantă cerință este că ele ar trebui să fie**
   a. colectate local de către instituțiile sau organizațiile care le vor utiliza
   b. foarte mari, numărând mii
   c. reprezentative pentru grupul pentru care vor fi utilizate
   d. obținute în mod convenabil în procesul de standardizare

5. **Conceptele de plafon al testului și prag al testului sunt cel mai strâns legate de aspectul**
   a. validității testului
   b. dificultății testului
   c. tipul de scoruri standard utilizate într-un test
   d. tipul de itemi utilizați într-un test

6. **Atunci când este transformat într-o scală IQ de tip Wechsler, un scor z de -1,00 va deveni un IQ Wechsler de**
   (a) 85.
   (b) 95.
   (c) 105.
   (d) 115.

7. **Care dintre următoarele procedee de transformare a scorurilor este singurul considerat drept transformare liniară?**
   (a) scoruri standard normalizate
   (b) percentile în stanine
   (c) scoruri brute în scoruri percentile
   (d) scoruri z în scoruri *T*

8. **Dacă unui grup de indivizi din ziua de azi ar trebui să i se aplice atât Scala Wechsler de inteligență pentru adulți inițială (WAIS), cât și ultima sa revizuire, WAIS — III, există șanse ca scorurile lor IQ să fie**
   a. aceleași la ambele teste
   b. mai mari la WAIS — III decât la WAIS
   c. mai mari la WAIS decât la WAIS-III

9. **Caracteristica esențială a modelelor teoriei răspunsului la item este că acestea plasează _____ pe o scală comună.**
   a. itemii și testele
   b. itemii și persoanele
   c. persoanele și testele

10. **Printre principalele avantaje ale metodei TRI se numără faptul că este ideală pentru utilizarea în testarea _____**
    a. computerizată adaptativă
    b. creion-hârtie
    c. personalității

Răspunsuri: 1. a; 2. d; 3. c; 4. c; 5. b; 6. a; 7. d; 8. c; 9. b; 10. a.

Aspecte esențiale ale fidelității

Termenul *fidelitate* sugerează încredere. În măsura în care, în întregime sau parțial, trebuie luate decizii pe baza scorurilor unui test, utilizatorii testului trebuie să fie siguri că pot avea o încredere destul de mare în scorurile obținute. Atunci când este folosită în legătură cu testele și măsurătorile, fidelitatea se bazează pe consistența și precizia rezultatelor procesului de măsurare. Pentru a avea un anumit grad de încredere în scoruri, utilizatorii testului cer dovezi ale faptului că scorurile obținute la test ar fi constante dacă testul ar fi aplicat în mod repetat aceloraşi indivizi sau grupuri și că scorurile au o precizie rezonabilă.

În timp ce fidelitatea în măsurare presupune consistență și precizie, lipsa fidelității presupune inconsistență și imprecizie, ambele echivalate cu o eroare de măsură. În contextul testării, *eroarea de măsură* poate fi definită ca orice fluctuație a scorurilor rezultată din factori legați de procesul de măsurare care sunt irelevanți pentru ceea ce este măsurat. Prin urmare, *fidelitatea* este o calitate a scorurilor obținute la test ce sugerează că acestea sunt suficient de consistente și de lipsite de erori de măsură pentru a fi utile.

Observați că, pentru a fi utile, scorurile obținute la un test nu trebuie să fie nici în totalitate consistente, nici lipsite de erori. Așa cum am văzut în Capitolul 1, chiar și în științele fizice — unele mândrindu-se cu instrumente incredibil de fidele — măsurătorile sunt întotdeauna subiectul unui anumit grad de eroare și fluctuație. În științele sociale și în științele comportamentului, măsurătorile sunt mult mai predispuse la erori datorită naturii evazive a constructelor care sunt evaluate și faptului că datele comportamentale prin care sunt evaluate pot fi influențate de mult mai mulți factori dificil de controlat decât alte tipuri de date (vezi Info 5.2 din Capitolul 5 despre Deconstruirea constructelor). În special scorurile testelor psihologice sunt deosebit de susceptibile la influențe provenite din diferite surse — printre care subiectul testat, examinatorul și contextul în care are loc testarea — toate putând determina o variabilitate exterioară dimensiunii măsurate de test.

Adevăr și eroare în măsurarea psihologică

Una dintre cele mai durabile abordări a temei fidelității este noțiunea de *scor real* (vezi Gullinksen, 1950) din teoria clasică a testării. Într-un fel, se poate spune că această noțiune reprezintă obiectul căutărilor sau „Graalul" psihometriei. Deși scorurile reale nu există cu adevărat, este posibil să ne imaginăm existența lor: scorurile reale sunt entități ipotetice care ar rezulta dintr-o măsurare lipsită de eroare. Metodele de estimare a fidelității scorurilor oferă un mod de a estima scorurile reale, sau, cel puțin, a limitelor în interiorul cărora s-ar putea situa scorurile reale. Conceptele de fidelitate și eroare a scorurilor testului — care, în mod evident, trebuie luate în considerare cu privire la orice scor — se aplică

în moduri asemănătoare, dar oarecum diferite, când este vorba despre scorurile obținute de un singur individ sau de un grup.

Conceptul de scor real în datele individuale

În teoria clasică a testării, *scorul real* al unui individ este conceptualizat ca scorul mediu al unei distribuții ipotetice de scoruri care s-ar obține dacă individului i s-ar aplica același test de un număr infinit de ori. În practică, este în mod evident imposibil să obții un astfel de scor, chiar și pentru un singur individ. În locul scorurilor reale, din test rezultă *scoruri observate* (i.e., scoruri pe care, de fapt, indivizii le obțin).

Cu privire la un singur scor, ideile prezentate până acum pot fi prezentate succint prin intermediul următoarei ecuații:

$$X_o = X_{real} + X_{eroare} \quad (4.1)$$

care exprimă ideea că orice scor observat (X_o) este alcătuit din două componente: componenta scorului real (X_{real}) și o componentă a scorului determinată de eroare (X_{eroare}). Dintr-un punct de vedere realist, mărimea celor două componente va rămâne întotdeauna necunoscută într-o situație dată. Cu toate acestea, în teorie, componenta scorului real este construită pentru a reprezenta acea parte a scorului observat care reflectă aptitudinea, trăsătura sau caracteristica pe care testul o evaluează. Invers, componenta erorii, definită ca diferența dintre scorul observat și scorul real, reprezintă alți factori care ar putea intra în scorul observat ca o consecință a procesului de măsurare.

Scoruri reale în datele de grup

Importanța singulară a variabilității inter-individuale a fost deja discutată în Capitolul 2, unde s-a arătat faptul că utilitatea testării psihologice depinde de obținerea unei anumite variabilități între indivizi. Fără o variabilitate a scorurilor, testele nu ne pot ajuta să luăm decizii cu privire la oameni bazate pe comparații. De asemenea, ne putem reaminti că, în același capitol, dispersia eșantionului (s^2) a fost definită ca gradul mediu de variabilitate într-un grup de scoruri. Pe baza acestei informații, Formula (4.1) — care se referă la un singur scor al testului — poate fi extrapolată și aplicată în următoarea manieră distribuției scorurilor la test obținute dintr-un eșantion sau o populație:

$$\text{Dispersia eșantionului} = s^2 = s_t^2 + s_e^2 \quad (4.2)$$

Sau

$$\text{Dispersia populației} = \sigma^2 = \sigma_t^2 + \sigma_e^2 \quad (4.3)$$

Ambele formule exprimă aceeași idee, și anume faptul că dispersia (varianța) unui set de scoruri observate pentru un eșantion (s^2) sau pentru o populație (σ^2) constă dintr-o parte care reprezintă dispersia (varianța) reală (s_t^2 sau σ_t^2) și o parte care reprezintă dispersia (varianța) erorii (s_e^2 sau σ_e^2). Dispersia reală constă din acele diferențe între scorurile indivizilor în cadrul unui grup care reflectă nivelul sau poziția lor la o anumită caracteristică pe care testul o evaluează. Dispersia erorilor este alcătuită din diferențele între scorurile testului care reflectă factori irelevanți pentru ceea ce evaluează testul. Formulele (4.2) și (4.3) implică și faptul că fidelitatea scorurilor crește pe măsură ce componenta

determinată de eroare scade. De fapt, un *coeficient de fidelitate* (r_{xx}) — despre care se vor spune mai multe mai târziu în acest capitol — poate fi definit ca raportul dintre dispersia reală a scorurilor (s_t^2) și dispersia totală a scorurilor testului (s^2) sau

$$r_{xx} = \frac{s_t^2}{s^2} \quad (4.4)$$

Cu alte cuvinte, dacă toată dispersia scorurilor testului ar fi o dispersie reală, fidelitatea scorurilor ar fi perfectă (1,00). Un coeficient de fidelitate poate fi văzut ca un număr care estimează proporția dispersiei dintr-un grup de scoruri obținute la test care este explicată prin erori ce provin din una sau mai multe surse. Din această perspectivă, evaluarea fidelității scorurilor presupune un proces în doi pași care constă din (a) determinarea posibilelor surse de eroare care pot interveni în scorurile testului și (b) estimarea mărimii acestor erori.

Relativitatea fidelității

Deși practica descrierii *testelor* ca fidele este obișnuită, de fapt, calitatea de fidelitate, dacă este prezentă, aparține nu testelor, ci *scorurilor* obținute la test. Această distincție este subliniată cu consecvență de cei mai mulți autori care contribuie la literatura de specialitate contemporană despre fidelitate (vezi Info 4.1). Chiar dacă poate părea subtilă la prima vedere, distincția este importantă pentru înțelegerea implicațiilor conceptului de fidelitate în utilizarea testelor și a interpretării scorurilor obținute la test. Dacă un test este descris ca fidel, implicația este că fidelitatea sa a fost stabilită permanent, sub toate aspectele, pentru toate utilizările și pentru toți utilizatorii. Ar fi asemănător cu a

spune că un pian care este bine acordat va fi acordat întotdeauna și va suna la fel de bine indiferent de tipul de muzică sau de persoana care cântă la el. De fapt, calitatea sunetului emis de un pian nu depinde doar de instrument, ci și de variabilele legate de muzică, de pianist și de cadrul în care se cântă la pian (de ex. acustica încăperii). Similar, deși fidelitatea în testare este legată într-o măsură semnificativă de caracteristicile testului, fidelitatea scorurilor testului — care reprezintă rezultatul utilizării instrumentului și, ca și muzica emisă de pian, este ceea ce contează cu adevărat — poate fi afectată de multe alte variabile.

Info 4.1

O excelentă colecție de lucrări despre fidelitate, cu explicații mai detaliate și extinse ale multor teme atinse în acest capitol poate fi găsită în *Score Reliability: Contemporary Thinking on Reliability Issues*, editată de Bruce Thompson (2003b).

Mai mult, chiar și atunci când este aplicată scorurilor obținute la test, calitatea de fidelitate este relativă. Scorul pe care o persoană îl obține la un test nu este fidel sau nefidel în sens absolut. Un scor obținut poate fi mai mult sau mai puțin fidel datorită factorilor care țin de persoana testată (oboseală, lipsa motivației, influența unor medicamente etc.) sau de condițiile situației de testare (prezența zgomotelor care distrag atenția, personalitatea examinatorului, strictețea cu care sunt impuse limitele de timp etc.). Toți acești factori, acționând independent sau combinat, pot influența scorul obținut într-o măsură mai mare sau mai mică, până la punctul în care scorul devine atât de lipsit de fidelitate, încât trebuie înlăturat. Chiar dacă nu au o legătură cu testul, toate aceste aspecte trebuie luate în considerare în procesul de evaluare.

Dimpotrivă, atunci când fidelitatea (r_{xx}) este luată în considerare din punctul de vedere al datelor scorurilor obținute *dintr-un*

eșantion mare, în condiții standardizate, se presupune că erorile de măsură care pot influența scorurile individuale ale membrilor eșantionului, deși încă prezente, sunt distribuite aleatoriu. Din moment ce erorile aleatorii au o probabilitate egală de a influența scorurile într-o direcție negativă sau pozitivă, se poate presupune că erorile se anulează una pe cealaltă. Cu toate acestea, chiar și în aceste condiții, estimările fidelității vor varia de la un eșantion la altul în funcție de componența lor și de circumstanțele în care are loc testarea. De exemplu, dacă un test este destinat adulților cu vârsta cuprinsă între 18 și 90 de ani, fidelitatea estimată a scorurilor sale va fi susceptibilă la influența unor factori diferiți, în funcție de datele pe care se bazează estimarea, dacă acestea sunt obținute din grupuri cu vârstă mai înaintată sau dintr-un grup reprezentativ pentru întregul interval de vârstă căruia îi este destinat testul.

Nu uitați!

- Fidelitatea este o caracteristică a *scorurilor* obținute la test, și nu a testelor.
- Fidelitatea oricărei măsurători și în special a scorurilor unui test psihologic, nu este nici absolută, nici imuabilă. Astfel, posibilele surse de eroare de măsură și gradul în care acestea sunt implicate într-o utilizare specifică a testului trebuie luate în calcul, estimate și raportate de fiecare dată când sunt utilizate scorurile obținute la test (AERA, APA, NCME, 1999).

O observație cu privire la adevăr și eroare

Nevoia de a identifica și investiga componentele reale și componentele erorii scorurilor este luată în considerare mai detaliat în următoarea secțiune. Totuși, este important să subliniem în acest punct că întotdeauna aceste judecăți trebuie să fie făcute în relație cu ceea ce testul încearcă să evalueze și cu circumstanțele în

care este administrat. De exemplu, scorurile unui test care denotă viteza cu care indivizii pot plasa 100 de cuie într-o placă, dacă este aplicat în condiții standardizate, ar putea reflecta, în mare, destul de fidel, nivelele tipice ale dexterității manuale ale celor supuși testării. Dacă același test este aplicat (a) în condiții menite să distragă atenția subiecților sau (b) în condiții care distrag neintenționat atenția subiecților, ambele seturi de scoruri vor reflecta nivele ale dexterității manuale în condiții de perturbare a atenției. Cu toate acestea, influența distragerilor va fi văzută drept sursă de varianță a erorilor sau ca reducând fidelitatea a ceea ce scorurile intenționează să indice doar în cea de-a doua situație.

Surse de eroare în testarea psihologică

Așa cum am văzut, eroarea poate intra în scorurile obținute la testele psihologice dintr-un număr enorm de motive, multe dintre acestea aflate în afara ariei estimărilor psihometrice ale fidelității. În general vorbind, totuși, erorile care apar în scorurile obținute la un test pot fi clasificate după următoarele trei surse: (a) contextul în care are loc testarea (incluzând factorii legați de persoana care administrează testul, de persoana care scorează testul și de motivele pentru care este realizată testarea), (b) persoana testată și (c) testul însuși. Unele dintre erorile care apar din aceste surse pot fi minimalizate sau eliminate, dat fiind faptul că practicile de testare adecvate sunt urmate de părți implicate în procesul de construire, selectare, administrare și scorare a testelor. Altele, cum ar fi lipsa de preocupare sau încercarea de a impresiona prin răspunsul la itemii testului nu pot fi eliminate, dar pot fi detectate de diferite tipuri de itemi de control incluși în test. Practicile legate de utilizarea adecvată a testelor — dintre care

multe au drept scop reducerea erorilor scorurilor — sunt discutate detaliat în Capitolul 7, care se ocupă de aspecte relevante pentru selectarea, administrarea și scorarea testului, printre altele.

În discutarea fidelității, se va presupune că utilizatorii testului, cei care îl administrează și îl scorează, selectează cu atenție instrumente adecvate, pregătesc medii de testare potrivite, stabilesc un raport bun cu subiecții testați și scorează testele în acord cu procedurile standardizate bine stabilite. Mai mult, se va presupune de asemenea că persoanele testate sunt pregătite în mod adecvat și au o motivație pentru a fi testate. Indiferent dacă aceste presupuneri sunt sau nu justificate în situații specifice, comportamentele pe care le presupun se află sub controlul unuia sau mai multor indivizi implicați în procesul de testare și nu țin în mod direct de test. În măsura în care aceste presupuneri sunt justificate, erorile din scorurile testului care apar din surse fără legătură cu testul pot fi, evident, eliminate sau, cel puțin, minimizate.

În continuare, în prezentarea fidelității, sursele de eroare ce vor fi discutate aparțin în principal primilor doi factori aflați în afara controlului conștient al părților implicate în procesul testării și anume factori întâmplători sau aleatori. Totuși, înainte de a trece mai departe, trebuie să observăm că eroarea de măsură poate fi sistematică și consistentă, dar și aleatorie. Așa cum o balanță poate greși cu câteva grame, tot așa un test poate avea o caracteristică intrinsecă de un anumit tip care afectează toți subiecții testați. Este posibil ca estimările tradiționale ale fidelității să nu reușească să detecteze acest tip de eroare consistentă datorită sursei acesteia, deoarece ele se bazează pe metode menite să detecteze inconsistențele din rezultatul unui test. Erorile sistematice și consistente de măsură afectează nu numai fidelitatea, dar și validitatea rezultatelor testului. Pentru a le detecta, trebuie să se compare

rezultatele unui instrument cu cele ale altor instrumente care evaluează același construct, dar nu au în comun factorul care cauzează eroarea consistentă. Pentru a detecta eroarea în cazul unei balanțe care măsoară cu câteva grame mai puțin, de exemplu, trebuie să cunoaștem greutatea acelei persoane sau a obiectului indicată de unul sau mai multe cântare bine calibrate.

Info 4.2 enumeră câteva dintre posibilele surse de eroare care pot face ca scorurile obținute la test să fie inconsistente. Această listă clasifică sursele de eroare evaluate de estimările tradiționale ale fidelității, tipurile de teste la care se referă acele surse de eroare în cel mai direct mod și coeficienții de fidelitate utilizați în mod tipic pentru a le estima. O explicație conceptuală a fiecărei surse de eroare și a fiecărei estimări a fidelității este prezentată în continuare, în aceeași ordine în care sunt enumerate în Info 4.2. Considerații cu privire la momentul și modul în care aceste concepte și proceduri sunt aplicate în procesul de utilizare a unui test sunt discutate în secțiunea următoare a acestui capitol.

Diferențele inter-evaluatori

Diferențele inter-evaluatori reprezintă o etichetă atribuită erorilor care pot intra în scoruri ori de câte ori elementul de subiectivitate joacă un rol în scorarea unui test. Se presupune că evaluatori diferiți nu vor atribui întotdeauna exact aceleași scoruri sau ranguri la un test de performanță dat, chiar dacă (a) instrucțiunile de scorare specificate în manualul testului sunt explicite și detaliate și (b) evaluatorii sunt conștiincioși în aplicarea acestor instrucțiuni. Cu alte cuvinte, variabilitatea scorurilor care se datorează diferențelor inter-evaluator nu presupune neatenție în pregătirea

Info 4.2

Surse ale erorii de măsură și coeficienții de fidelitate tipici utilizați pentru a le estima

Surse de eroare	Tipuri de teste predispuse la fiecare sursă de eroare	Măsuri adecvate utilizate pentru a estima eroarea
Diferențele dintre evaluatori	Teste scorate cu un grad de subiectivitate	Fidelitatea inter-evaluatori
Eroarea de eșantionare a timpului	Teste ale trăsăturilor sau comportamentelor relativ stabile	Fidelitatea test-retest (r_{tt}), numită și coeficient de stabilitate
Eroarea de eșantionare a conținutului	Teste pentru care este dorită consistența rezultatelor, ca întreg	Fidelitatea estimată prin forme alternative (r_{ll}) sau prin metoda înjumătățirii
Inconsistența inter-item	Teste care cer consistența inter-item	Fidelitate estimată prin metoda înjumătățirii sau măsuri ale consistenței interne mai precise, cum ar fi coeficientul Kuder-Richardson 20 (K-R 20) sau coeficientul alfa (α).
Inconsistența inter-item combinată cu eterogenitatea conținutului	Teste care cer consistența inter-item și omogenitate	Măsuri ale consistenței interne *și* dovezi suplimentare ale omogenității
Erori combinate de eșantionare a timpului și conținutului	Teste care cer stabilitatea *și* consistența rezultatelor ca întreg	Fidelitate estimată prin forme paralele aplicate după un interval mai mare de timp

instrucțiunilor de scorare și în scorarea concretă a unui test. Se referă la variații ale scorurilor care apar datorită diferențelor dintre judecățile subiective ale evaluatorilor.

Fidelitatea inter-evaluatori

Metoda de bază pentru estimarea erorii datorate diferențelor dintre evaluatori constă în scorarea de către cel puțin doi indivizi diferiți a aceluiași set de teste, astfel încât sunt generate două sau mai multe scoruri independente pentru performanța fiecărei persoane testate. Corelațiile dintre seturile de scoruri generate în această manieră sunt indicatori ai *fidelității inter-evaluatori*. Corelații foarte puternice și pozitive, de ordinul 0,9 sau mai înalte, sugerează că proporția de eroare explicată prin diferențele inter-evaluatori este de 10 % sau mai mică, deoarece 1 – (> 0,90) = < 0,10.

Eroarea determinată de eșantionarea timpului

Eroarea determinată de eșantionarea timpului se referă la variabilitatea inerentă în scorurile testului datorită faptului că acestea sunt obținute la un anumit moment de timp, și nu în altul. Acest concept este legat de două concepte relaționate, și anume (a) că este probabil ca orice construct sau comportament pe care un test îl evaluează să fluctueze în timp și (b) că unele dintre constructele și comportamentele evaluate prin teste fie sunt mai puțin supuse schimbării, fie se schimbă într-un ritm mai lent decât altele. De exemplu, constructele psihologice legate de aptitudini, cum ar fi înțelegerea verbală sau aptitudini mecanice, sunt văzute de obicei ca fiind mai puțin predispuse la fluctuații în timp decât constructele legate de personalitate, cum ar fi agreabilitatea sau căldura afectivă. În domeniul personalității și al funcționării emoționale,

diferența dintre constructe mai mult sau mai puțin stabile a fost codificată prin distincția tradițională între *trăsături* — care sunt considerate drept caracteristici relativ durabile — și *stări*, care, prin definiție, sunt condiții temporare. Într-o anumită măsură, această distincție poate fi aplicată și caracteristicilor cognitive. De exemplu, se presupune că aptitudinea verbală este mult mai stabilă în cazul unui individ decât capacitățile sale de atenție și memorie, acestea fiind mai susceptibile la influența condițiilor trecătoare sau a stărilor emoționale. În orice caz, este clar că, din moment ce se presupune că un anumit grad de eroare determinată de eșantionarea timpului intră în toate scorurile testului, ca regulă, ne putem aștepta ca o parte mai mică să intre în scorurile testului care evaluează trăsături relativ stabile.

Fidelitatea test-retest

Pentru a genera estimări ale gradului de eroare determinată de eșantionarea timpului care are probabilitatea de a influența scorurile unui test dat, se administrează de obicei același test în două situații diferite, separate de un anumit interval de timp, unuia sau mai multor grupuri de indivizi. Corelația dintre scorurile obținute în cele două administrări reprezintă *coeficientul de fidelitate test-retest* r_{tt} (sau *de stabilitate*) și poate fi văzut ca un indicator al măsurii în care este probabil ca scorurile să fluctueze ca rezultat al erorii de eșantionare a timpului. Atunci când este utilizată această procedură, intervalul de timp dintre cele două administrări ale testului trebuie specificat întotdeauna, deoarece va afecta în mod evident stabilitatea scorurilor. Realist vorbind, totuși, există mulți factori care pot influența diferențiat scorurile obținute de la un grup de oameni în două situații. Din acest motiv, nu există un interval de timp fix, care poate fi recomandat pentru

orice test. Dacă intervalul este foarte redus, de exemplu, cei cărora li se aplică testul își pot aminti răspunsurile din prima situație și aceasta le poate influența scorurile obținute la retestare. Pe de altă parte, dacă intervalul este prea lung, există întotdeauna posibilitatea ca experiențele care intervin — inclusiv pașii pe care subiectul testat i-ar fi putut face ca reacție la prima administrare — să poată afecta scorurile obținute în a doua situație. Pe lângă aceste aspecte, utilizatorii testului pot evalua și coeficienții de stabilitate din punctul de vedere al așteptărilor teoretice cu privire la trăsăturile și comportamentele evaluate de test. Un exemplu ar putea fi diferențele apărute în rata schimbării care pot fi așteptate în funcție de vârsta celor cărora li se aplică testul. Înțelegerea textului citit, de exemplu, se poate schimba destul de rapid la copii, dar va rămâne stabilă la vârsta adultă, în afară de unele circumstanțe neobișnuite care o influențează — cum ar fi o formare specială sau leziuni cerebrale.

Eroarea determinată de eșantionarea conținutului

Eroarea determinată de eșantionarea conținutului este un termen utilizat pentru a eticheta variabilitatea irelevantă pentru trăsătură care poate intra în scorurile testului ca rezultat al factorilor fortuiți legați de conținutul itemilor specifici incluși într-un test. Un exemplu simplu al modului în care eroarea de eșantionare a conținutului ar putea intra în scorurile testului este prezentat în Info 4.3. Această ilustrare a erorii determinată de eșantionarea conținutului prezintă un caz ipotetic, deoarece este vorba despre o eroare apărută în scorurile obținute la test ca rezultat al unei construiri defectuoase a testului care ar fi putut fi evitată cu ușurință. În exemplul nostru, modul în care profesorul selectează

Info 4.3

O ilustrare simplă a erorii de eșantionare a conținutului care rezultă din construirea greșită a testului

Să luăm în considerare cazul unui test educațional cu raportare la conținut, destinat să evalueze cunoașterea *întregului* material conținut în cinci capitole ale unui manual. Să presupunem că profesorul care a pregătit testul a construit cea mai mare parte a itemilor pornind numai de la conținutul a trei capitole, neglijând să includă itemi din cele două capitole rămase. Să presupunem mai departe că unii elevi s-au concentrat de asemenea doar pe trei capitole în studiul lor în vederea testării.

Eroarea determinată de eșantionarea conținutului apărută în scorurile la un astfel de test va rezulta în primul rând din eșantionarea inegală a materialului pe care testul trebuie să îl acopere. *Toate celelalte aspecte rămânând neschimbate,* consecințele eșantionării inadecvate a conținutului de către profesor vor fi următoarele: (a) acei elevi care au studiat aceleași capitole din care a fost extras conținutul testului vor scora aproape 100%, (b) cei care s-au concentrat pe două din acele capitole și unul dintre celelalte vor obține note în jur de 67% și (c) cei care au avut nenorocul de a se concentra numai pe unul dintre capitolele „corecte" și pe cele două dintre capitolele neincluse în test, vor obține un scor de aproximativ 33%.

Dacă presupunem că toți studenții care au studiat doar trei capitole din cinci au stăpânit de fapt 60% din materialul pe care testul trebuia să îl acopere, scorurile lor reale ar fi trebuit să aproximeze acest procent. Discrepanțele dintre scorurile obținute și nivelul lor real de stăpânire a materialului reprezintă eroarea de eșantionare a conținutului. În acest caz particular, eroarea introdusă în scoruri va scădea nu numai fidelitatea acestora, dar și validitatea lor.

itemii duce la o acoperire neadecvată a conținutului de cunoștințe pe care se presupune că testul le evaluează. Ca urmare, o bună parte din variabilitatea scorurilor nu este legată de nivelul de asimilare de către studenți a materialului specificat și face ca

scorurile să fie nu doar mai puțin fidele, dar și mai puțin valide decât în alt caz. Un exemplu tipic este oferit de cazurile în care — din motive aflate în afara controlului persoanei care construiește testul — conținutul specific al unui test fie îi avantajează, fie îi dezavantajează pe unii dintre indivizii testați, în funcție de experiențele lor anterioare. De exemplu, un test construit pentru a evalua înțelegerea unui text poate include în mod fortuit câteva pasaje familiare anumitor subiecți și nefamiliare altora. În mod evident, cei care sunt familiarizați cu pasajele vor fi capabili să răspundă la întrebări pe baza lor mult mai repede și cu mai mare ușurință datorită familiarizării lor mai mari cu materialul, și nu datorită unui nivel mai înalt al achizițiilor din sfera înțelegerii.

Estimarea fidelității prin metoda formelor alternative

Procedeele *fidelității estimate prin forme alternative* sunt destinate să estimeze gradul de eroare din scorurile obținute la test care poate fi atribuit erorii de eșantionare a conținutului. Pentru a investiga acest tip de fidelitate, două sau mai multe forme diferite ale testului — identice ca scop, dar diferite sub aspectul conținutului specific — trebuie pregătite și administrate aceluiași grup de subiecți. Scorurile pentru fiecare versiune sunt apoi corelate pentru a obține *coeficienții de fidelitate prin forme alternative* (r_{11}). Deoarece este improbabil ca aceiași factori determinați de întâmplare, care favorizează anumiți subiecți și nu alții, să influențeze formele alternative, corelațiile puternice și pozitive (ex.: 0,90 sau mai mari) între scorurile la diferite forme pot fi luați ca un indicator al faptului că eroarea de eșantionare a conținutului nu are o influența majoră asupra scorurilor obținute la test (ex.: 10% sau mai puțin). Fidelitatea prin forme alternative decalate în timp, o variantă a acestei proceduri utilizată pentru a evalua efectele

Nu uitați!

Formularea *toate celelalte aspecte rămânând neschimbate* apare în Info 4.3 și în altă parte în această carte. Această frază este o formulă retorică, menită să sugereze că toate considerațiile pertinente, altele decât conceptele luate în discuție în acel moment, trebuie trecute cu vederea temporar pentru izolarea și clarificarea aspectelor aflate în discuție. Totuși, merită amintit că în testare, ca și în alte aspecte ale vieții, la o privire realistă, „toate celelalte aspecte" rămân neschimbate foarte rar, dacă se întâmplă astfel vreodată. Formularea va servi (a) pentru a-l avertiza pe cititor de posibilitatea ca alte câteva aspecte să fie luate în considerare, pe lângă conceptul specific aflat în discuție și (b) pentru a-l stimula pe cititor în a lua în considerare cu atenție celelalte aspecte care ar putea exista.

combinate ale eșantionării timpului și conținutului, este discutată mai târziu în această secțiune.

Estimarea fidelității prin metoda înjumătățirii

Construirea formelor alternative ale unui test sau administrarea aceluiași test de două ori presupune deseori probleme teoretice și practice care fac dificil acest demers. O soluție este să se administreze pur și simplu un test unui grup de indivizi și să se creeze două scoruri pentru fiecare persoană, împărțind testul pe jumătate, în funcție de modul în care acesta este construit, pentru a calcula coeficienții de fidelitate prin înjumătățire. În particular, este imperativ să se ia în considerare două posibilități: (a) dacă anumiți itemi ai testului diferă în mod sistematic de ceilalți itemi din test și (b) dacă viteza joacă un rol semnificativ în performanța la test. Ambele condiții pot avea efecte profunde asupra mărimii coeficienților de fidelitate prin înjumătățire.

1. *Diferențe sistematice ale itemilor testului* pot apărea din diferite motive. De exemplu, multe teste de aptitudini încep cu itemii cei mai ușori și devin progresiv din ce în ce mai dificile sau sunt divizate în părți și subteste care acoperă conținuturi diferite. Altele, ca Wonderlic Personnel Test, sunt structurate într-un *format spirală* astfel încât itemii care au de-a face cu sarcini verbale, numerice, spațiale și analitice sunt rotați sistematic. De asemenea, numeroase inventare de personalitate sunt aranjate astfel încât itemi din scale diferite apar prin rotație de-a lungul testului.
2. *Atunci când performanțele la test depind în primul rând de viteză*, itemii sunt de obicei fixați la un nivel de dificultate suficient de scăzut pentru toți subiecții testați, astfel încât toți îi pot rezolva corect, dar limitele de timp sunt stabilite astfel încât cei mai mulți dintre ei nu vor fi capabili să termine testul. De exemplu, testele de abilități funcționărești includ deseori sarcini care cer ca persoanele testate să urmărească atent o listă lungă de perechi de numere, litere sau simboluri pentru o scurtă perioadă de timp și să indice dacă fiecare pereche este sau nu identică. În acest tip de test cu viteză foarte mare, scorurile depind în principal de numărul de itemi completați, și nu de numărul de răspunsuri corecte. Deoarece cea mai mare parte a persoanelor testate vor obține rezultate perfecte sau aproape perfecte la toți itemii pe care îi încearcă, orice împărțire a unui astfel de test în jumătăți — ca și orice măsură a consistenței interne — va duce la coeficienți aproape perfecți.

Info 4.4 prezintă posibile soluții ale problemei modului în care diferite tipuri de teste pot fi înjumătățite. Odată ce acest lucru a fost realizat, este utilizată corelația dintre scorurile obținute la una dintre jumătățile testului și scorurile la cealaltă jumătate (r_{hh})

Info 4.4

Câteva soluții ale problemei împărțirii unui test în jumătate

- O regulă practică a împărțirii testelor de diferite tipuri pentru a calcula coeficienți de fidelitate prin înjumătățire este de a *diviza testul în două jumătăți comparabile.* Deși aceasta se poate realiza în multe moduri, deseori se utilizează o metodă de distribuire par-impar, cu itemii impari (1, 3, 5 etc.) și itemii pari (2, 4, 6 etc.) alcătuind cele două jumătăți.
- Dacă viteza joacă un rol în performanța la test, orice estimare a fidelității într-o singură probă — cum ar fi metoda înjumătățirii –va produce în mod fals rezultate înalte. Aceasta se întâmplă deoarece, pentru testele în care viteza este semnificativă, fidelitatea scorurilor depinde în primul rând de consistența vitezei cu care subiecții realizează proba, opusă consistenței răspunsurilor lor. Astfel, pentru testele de viteză, o posibilă soluție este utilizarea *metodelor de estimare a fidelității prin două probe,* cum ar fi test-retest sau formele alternative. O altă posibilitate este aceea de a împărți testul în jumătăți cronometrate separat, urmată de calcularea coeficientului de fidelitate în aceeași manieră ca în cazul obișnuit al metodei înjumătățirii.
- *De ce este acest fapt important pentru un potențial utilizator?* Dacă metoda utilizată pentru a calcula estimări ale consistenței interne a scorurilor testului nu este adecvată modului în care un test a fost construit, coeficienții de fidelitate rezultați ar putea fi înșelători. Utilizatorii potențiali care iau în considerare aspectul fidelității în procesul de selectare a unui test trebuie să acorde atenție acestor probleme.

pentru a obține un *coeficient de fidelitate prin înjumătățire.* Din moment ce r_{hh} estimează consistența scorurilor în cele două jumătăți ale testului, *formula Spearman-Brown* (S-B) este aplicată lui r_{hh} pentru a obține o estimare pentru întregul test. Această formulă se bazează pe ideea din teoria clasică a testării conform căreia un număr mare de observații va produce un rezultat cu o

fidelitate mai mare decât un număr mai redus de observații. Cu alte cuvinte, un scor obținut la un test mai lung va fi mai apropiat de scorul real decât un scor obținut la un test mai scurt. Versiunea generală a formulei S-B este

$$r_{S-B} = \frac{nr_{xx}}{1 + (n-1)\, r_{xx}} \tag{4.5}$$

unde

r_{S-B} = estimarea Spearman-Brown a coeficientului de fidelitate

n = multiplul cu care lungimea testului trebuie crescută sau scăzută

r_{xx} = coeficientul de fidelitate obținut pentru lungimea inițială a testului

Această formulă poate fi utilizată pentru a estima efectul pe care îl va avea asupra coeficientului obținut extinderea lungimii unui test în orice grad sau scurtarea unui test prin orice fracție a dimensiunii sale inițiale. De exemplu, dacă r_{xx} pentru un test alcătuit din 30 de itemi este 0,80 și dorim să estimăm care ar fi coeficientul de fidelitate dacă lungimea testului ar crește la 90 de itemi, adăugându-se 60 de itemi *comparabili, n* ar fi egal cu 3 și r_{S-B} ar fi 0,92. Dacă dorim să scurtăm testul la 15 itemi, n va fi 1/2 și r_{S-B} ar fi sub 0,67. Atunci când formula S-B este aplicată unui coeficient de fidelitate prin înjumătățire (r_{hh}) care presupune estimarea fidelității întregului test bazat pe corelația dintre cele două jumătăți ale sale, ea poate fi simplificată astfel:

$$r_{S-B} = \frac{2r_{hh}}{1 + r_{hh}} \tag{4.6}$$

Inconsistența inter-item

Inconsistența inter-item se referă la erorile scorurilor care rezultă din fluctuațiile *itemilor* în întregul test, opusă erorii de eșantionare a conținutului care rezultă din configurarea particulară a itemilor incluși în întregul test. Deși inconsistențele inter-item pot fi evidente la o examinare atentă a conținutului de itemi — și a proceselor cognitive care pot intra în joc în răspunsul la diferiți itemi dintr-un test — din punct de vedere statistic acestea se manifestă prin corelații reduse între itemii testului. Astfel de inconsistențe se pot datora unor factori diferiți, printre care eroarea de eșantionare a conținutului, mulți dintre aceștia fiind fortuiți și impredictibili. Ele pot rezulta și din eterogenitatea conținutului.

Eterogenitatea conținutului

Eterogenitatea conținutului apare din includerea unor itemi sau seturi de itemi care se referă la cunoașterea conținutului sau la funcții psihice ce diferă de cele vizate de alți itemi din același test. Acest factor este în mare măsură controlat în etapa de construire a testului, când ar trebui să se determine gradul de eterogenitate al conținutului testului pe baza scopului testului și a tipului de populație căreia îi este adresat testul. În măsura în care un test este construit intenționat pentru a eșantiona un conținut eterogen, eterogenitatea conținutului nu poate fi considerată o sursă de eroare. Eterogenitatea conținutului testului sau a funcțiilor cognitive atinse de diferiți itemi ai testului este o sursă de eroare numai atunci când se presupune că un test este omogen în privința itemilor săi în unul sau mai multe moduri. Info 4.5 arată câteva seturi de itemi care variază sub aspectul eterogenității lor.

Nu uitați!

Analiza caracterului eterogen și omogen

În testarea psihologică, conceptele de eterogenitate și omogenitate sunt utilizate în raport cu componența (a) eșantioanelor de comportament sau a itemilor ce alcătuiesc un test și (b) grupurilor de persoane cărora li se aplică testul, cum sunt eșantioanele de standardizare sau populațiile. Deoarece ambele aspecte pot influența *toate* statisticile utilizate pentru a evalua testele (vezi secțiunea despre restricțiile de interval și despre corelație, precum și Fig. 2.7 din Capitolul 2), este important să reținem următoarele:

- *Eterogenitatea și omogenitatea sunt întotdeauna termeni relativi.* Orice entitate alcătuită din elemente separate este eterogenă în măsura în care elementele sale nu sunt similare sub anumite aspecte. Astfel, orice grup alcătuit din entități multidimensionale, oameni sau itemi ai testului, este eterogen din anumite puncte de vedere. Din același motiv, niciun astfel de grup nu este omogen sub *toate* aspectele.
- *Pentru a caracteriza un grup drept eterogen sau omogen,* este necesar să decidem care variabilă sau variabile vor servi ca bază pentru evaluarea similarității sau nesimilarității. De exemplu:
- *Itemii dintr-un test* pot fi eterogeni în raport cu conținutul și formatul — dacă unii sunt alcătuiți din cuvinte în timp ce alții sunt alcătuiți din numere sau unii sunt prezentați oral, alții în scris, dar pot fi omogeni în raport cu funcția cognitivă investigată, dacă toți implică memoria (e.g. amintirea cuvintelor sau a numerelor).
- *Un grup de persoane* poate fi eterogen în raport cu sexul și vârsta, dacă include bărbați și femei cu vârsta cuprinsă între 17 și 45 de ani, dar omogen în raport cu statutul educațional, dacă include numai studenți în anul întâi.

Măsuri ale consistenței interne

Măsurile consistenței interne sunt procedee statistice destinate evaluării inconsistenței itemilor testului. Coeficienții de fidelitate prin înjumătățire îndeplinesc această condiție într-o anumită măsură. Cu toate acestea, chiar și un test foarte scurt poate fi împărțit în jumătăți în câteva moduri diferite — de exemplu, un test alcătuit din patru itemi poate fi înjumătățit în 3 moduri diferite, un test alcătuit din șase itemi în 10 moduri și așa mai departe — și fiecare înjumătățire poate duce la o corelație oarecum diferită între cele două jumătăți. O cale de a depăși această problemă logistică este de a realiza o distribuire par-impar — cu o jumătate a testului alcătuită din itemi pari și cealaltă jumătate constând din itemi impari — sau orice altă metodă prin care ar rezulta două jumătăți comparabile (vezi Info 4.4).

O altă soluție este dată de formule care țin cont de corelația inter-item (corelația dintre performanțele la *toți itemii* din cadrul unui test). Cele două formule cel mai frecvent utilizate pentru a calcula consistența inter-item sunt *formula Kuder-Richardson 20* (*K-R 20*) și *coeficientul alfa* (α), cunoscut și ca *alfa Cronbach* (Cronbach, 1951) care este pur și simplu un caz mai general al postulatului K-R 20. Atât K-R 20, cât și coeficientul alfa necesită o singură administrare (o probă) a testului la un grup de indivizi. Mărimea coeficientului K-R 20 și a coeficienților alfa depinde de doi factori: (a) numărul de itemi din test și (b) raportul dintre variabilitatea performanțelor persoanelor testate la itemii testului și dispersia totală a scorurilor testului. Toate celelalte aspecte rămânând neschimbate, mărimile coeficientului K-R 20 și ale coeficientului alfa vor fi mai înalte (a) pe măsură ce numărul de itemi crește și (b) pe măsură ce raportul dintre dispersia scorului unui item și dispersia totală a scorurilor testului descrește. Conceptual, atât

Info 4.5

Exemple de seturi de itemi ai testului de la conținutul cel mai eterogen la conținutul cel mai puțin eterogen

Set (A)

Itemul 1: Ce număr ar trebui să urmeze în următoarea serie?
3 6 12 24 _____

Itemul 2: Care dintre următoarele cinci elemente enumerate este cel mai puțin asemănător cu celelalte patru?
Porc vită pui ton viței

Itemul 3: Un tren se deplasează 12 m în 1/2 secundă. La aceeași viteză, ce distanță va parcurge în 4 secunde?

Set (B)

Itemul 1: 4 + 10 = _____

Itemul 2: Dacă o duzină de ouă costă 1,03$, cât vor costa trei duzini de ouă?

Itemul 3: Prețul unui exemplar este redus cu 60% în timpul unei perioade de reduceri de prețuri. Cu ce procent ar trebui crescut prețul pentru a ajunge din nou la prețul inițial?
60% 80% 100% 120% 150%

Set (C)

Itemul 1: 4 x 5 = _____

Itemul 2: 7 x 11 = _____

Itemul 3: 15 x 15 = _____

- *Setul A este cel mai eterogen:* itemii diferă în termenii domeniilor de conținut, formatului și abilităților necesare.
- *Următorul este setul B:* Itemii aparțin aceluiași domeniu de conținut (matematică), dar diferă ca format și abilități necesare (i.e. adunare, înmulțire și fracții *plus* abilități fundamentale de citire).
- *Setul C este cel mai omogen:* Itemii săi au domeniu, format și abilități necesare comune (înțelegerea operației de înmulțire și a simbolurilor sale)

coeficientul K-R 20, cât și coeficientul alfa produc estimări ale fidelității echivalente cu media tuturor coeficienților de înjumătățire posibili care ar rezulta din toate modurile diferite posibile de înjumătățire a testului. Astfel, ele reprezintă o estimare combinată a erorii de eșantionare a conținutului și a eterogenității conținutului. Prin urmare, în afară de cazul în care un test este foarte omogen, fidelitatea K-R 20 și coeficientul alfa vor fi mai scăzute decât un coeficient de înjumătățire. Info 4.6 conține formula coeficientului K-R 20 și o versiune a formulei coeficientului alfa, alături de explicația de bază a componentelor și aplicabilității lor.

Din moment ce atât coeficientul K-R 20, cât și coeficientul alfa, depind puternic de gradul de variabilitate inter-item dintr-un test, rezultă că orice lipsă de uniformitate a conținutului itemilor testului, cum ar fi eterogenitatea conținutului, va duce la valori mai scătute ale acestor coeficienți. De exemplu, să presupunem că au fost calculați coeficienții consistenței interne pentru trei teste de lungime egală — alcătuite din itemi ca cei prezentați în seturile A, B și C din Info 4.5. Dacă acest lucru a fost făcut, testul care se aseamănă cu ansamblul de itemi din setul A în termenii eterogenității va avea cea mai scăzută consistență internă deoarece diferențele în stăpânirea diferitelor abilități de către cei cărora li s-a aplicat testul și domeniile de conținut atinse de itemii testului se vor reflecta în performanțele lor. Testul cu itemii cei mai omogeni, și anume itemi cum ar fi cei din setul C, vor avea cel mai înalt coeficient.

De ce este important acest fapt? Atunci când un test este proiectat intenționat pentru a include itemi diverși în raport cu una sau mai multe dimensiuni, coeficientul K-R 20 și coeficientul alfa vor supraestima eroarea de eșantionare a conținutului și, ca urmare,

Info 4.6

Formule pentru calcularea consistenței interne

Formula Kuder-Richardson 20 (K-R 20)

$$r_{\text{K-R 20}} = \left(\frac{n}{n-1}\right)\frac{s_t^2 - \Sigma pq}{s_t^2}$$

Unde

n = numărul de itemi ai testului

s_t^2 = dispersia scorului total al testului

Σpq = suma produselor pq pentru fiecare item al testului

p = proporția de persoane care au rezolvat corect fiecare item sau au răspuns la el într-un anumit fel

q = proporția de persoane care au greșit la fiecare item sau au dat un răspuns în direcția opusă.

Formula $r_{\text{K-R 20}}$ se aplică testelor ai căror itemi sunt scorați drept corect sau greșit sau în orice altă manieră dihotomică, cum ar fi adevărat sau fals *dacă* toți itemii sunt formulați astfel încât semnificația fiecărei alternative este uniformă în întregul test.

Coeficientul alfa (α) sau alfa Cronbach

$$\alpha = \left(\frac{n}{n-1}\right)\frac{s_t^2 - \Sigma\left(s_i^2\right)}{s_t^2}$$

unde

n = numărul de itemi ai testului

s_t^2 = dispersia scorurilor totale la test

$\Sigma\left(s_i^2\right)$ = suma dispersiilor scorurilor la item

Aceste formule, a coeficientului alfa și a variantei cunoscute drept *coeficient alfa standardizat*, care utilizează corelația medie inter-item în locul scorului itemului și a dispersiei scorului total, sunt utilizate pentru teste ai căror itemi au mai multe răspunsuri posibile (*acord puternic, acord, dezacord, și dezacord puternic*). Cortina (1993) discută pe larg semnificația formulelor coeficientului alfa și diferiți factori care pot influența rezultatele lor.

sunt neadecvați. În funcție de construirea testului și pe baza unei examinări a conținutului său, itemii omogeni pot fi plasați în subteste sau pot fi separați altfel pentru a calcula măsuri separate ale consistenței inter-item în grupurile cu itemi similari. Pe de altă parte, atunci când este urmărită omogenitatea tuturor itemilor testului, mărimile coeficienților K-R 20 și alfa reprezintă indicatori ai măsurii în care acest obiectiv a fost atins. De fapt, diferența dintre mărimea celui mai adecvat coeficient de fidelitate prin înjumătățire și mărimea coeficientului K-R 20 sau a coeficientului α poate fi considerată un indicator al gradului de eterogenitate a itemilor dintr-un test. Cu cât sunt mai apropiate cele două estimări, cu atât conținutul testului este mai eterogen. De asemenea, pot fi utilizate tehnici de analiză factorială pentru a investiga eterogenitatea și multidimensionalitatea itemilor testului. Aceste tehnici, discutate pe larg în Capitolul 5, sunt utilizate pentru a detecta similaritățile dintr-un set de variabile — cum ar fi răspunsurile la itemii testului — pe baza interrelaționării tiparelor lor de variabilitate în unul sau mai multe grupuri de subiecți testați.

Erori combinate determinate de eșantionarea timpului și a conținutului

Erorile determinate de eșantionarea timpului și a conținutului pot fi estimate într-o manieră combinată pentru testele care necesită atât stabilitate, cât și consistență a rezultatelor. Așa cum vom vedea pe scurt, este de asemenea posibil să se estimeze efectele combinate ale altor surse de eroare ale scorurilor testelor prin alte mijloace. Cu toate acestea, modelul formelor alternative decalate în timp oferă o bună metodă de estimare a erorii de eșantionare a timpului și a conținutului cu un singur coeficient.

Fidelitatea estimată prin forme alternative decalate în timp

Coeficienții *de fidelitate estimată prin forme decalate în timp* pot fi calculați atunci când două sau mai multe forme alternative ale aceluiași test sunt administrate în două situații diferite, separate de un anumit interval de timp, unuia sau mai multor grupuri de indivizi. Ca și în cazul fidelității test-retest, intervalul dintre cele două administrări trebuie să fie specificat în mod clar, alături de componența eșantioanelor și a altor condiții care ar putea influența mărimea coeficienților obținuți. Dacă cele două forme ale unui test sunt administrate în succesiune imediată sau apropiată în timp, coeficientul formelor alternative care rezultă va depinde în primul rând de fidelitatea celor două forme. În cazul unor intervale mai lungi între administrările celor două forme, varianța erorii scorurilor va reflecta fluctuațiile în timp și eroarea de eșantionare a conținutului prezente în test.

O notă cu privire la efectele practice. O consecință inevitabilă a utilizării aceluiași test sau a formelor alternative ale unui test, în mod repetat cu aceiași subiecți este că introduce în scoruri o sursă suplimentară de variabilitate nedorită datorită *efectelor exersării*. Bineînțeles, lungimea intervalului dintre administrări influențează măsura în care scorurile la a doua administrare sau la administrările ulterioare ale testului sunt supuse efectului exersării. Pentru intervale reduse de timp, efectele exersării pot fi destul de semnificative, mai ales atunci când itemii testului presupun noi sarcini care necesită ca anumiți subiecți să utilizeze anumite strategii de rezolvare de probleme care, probabil, pot fi reamintite. Metodele care utilizează o probă unică pentru estimarea fidelității scorurilor, cum ar fi metoda înjumătățirii și coeficientul alfa, nu sunt predispuse la efectele exersării, spre deosebire de metodele

pentru care sunt necesare două probe, ca fidelitatea test-retest și prin forme alternative, care, de obicei, sunt supuse acestor influențe. În măsura în care indivizii diferă în privința gradului de îmbunătățire a rezultatelor la retestare ca efect al exersării, este probabil ca și corelațiile obținute între cele două probe să fie reduse. Mai important totuși este faptul că atunci când testele sunt administrate în mod repetat pentru evaluarea schimbărilor apărute de-a lungul timpului, așa cum se întâmplă în studiile longitudinale, efectul exersării poate constitui o variabilă semnificativă care produce confuzie și de care trebuie să se țină cont. (vezi Kaufman & Lichtenberger, 2002, p. 163–165).

Fidelitatea în utilizarea testului

Fidelitatea scorurilor este o preocupare continuă în testarea psihologică datorită posibilității prezente permanent ca erorile apărute din diferite surse să intre în rezultatele testului. Cu toate acestea, maniera în care fidelitatea scorurilor este luată în considerare variază în diferite momente ale procesului de construire a unui test și în aplicarea concretă a testelor. Din perspectiva utilizatorului testelor, estimările fidelității trebuie să fie luate în considerare cu atenție și aplicate în stadiile de (a) selectare a testului și (b) interpretare a scorurilor obținute la test. Capitolul 7 se ocupă în detaliu de aceste probleme; cu toate acestea, din moment ce utilizările estimărilor statistice ale fidelității sunt prezentate în acest capitol, modurile diferite în care fidelitatea este luată în considerare în fiecare stadiu vor fi prezentate în acest punct pentru a oferi un context pentru discuția ulterioară.

Considerații cu privire la utilizarea fidelității în selectarea testelor

Atunci când utilizatorii se decid ce test să folosească într-un scop dat, ei trebuie să cerceteze datele existente cu privire la fidelitatea scorurilor obținute la un anumit test. De obicei, aceste date pot fi găsite în manualele testelor, în cărți și în articole pregătite de autorii testului, dar pot apărea și în literatura de specialitate ca rezultat al activității unor cercetători independenți care au utilizat acele teste. În mod tipic, datele despre fidelitate sunt prezentate sub forma coeficienților de corelație. Datorită utilizării generalizate a coeficientului Pearson *r* în evaluarea fidelității și validității scorurilor la test, aspectele esențiale ale acestei metode de corelație, inclusiv limitele sale (discutate în Capitolul 2) trebuie înțelese complet înainte de a avansa în acest subiect. Un fapt deosebit de relevant care trebuie reținut este că mărimea coeficienților de corelație depinde într-o anumită măsură de variabilitatea eșantioanelor pentru care au fost calculate (vezi secțiunea despre Restricții de interval și corelații din Capitolul 2).

Diferitele tipuri de coeficienți care pot fi calculați pentru a estima eroarea de măsură, alături de cele mai relevante surse de eroare, au fost deja descrise, deși în mod abstract. În momentul selectării unui test, potențialul utilizator al testului trebuie să aplice aceste noțiuni la situațiile particulare pentru care doresc să utilizeze un test. Info 4.7 enumeră pașii fundamentali ai selectării unui test din punctul de vedere al fidelității; o discuție mult mai detaliată a acestor considerații este prezentată în următoarele paragrafe. Ca avertisment preliminar, trebuie notat că în evaluarea caracteristicilor psihometrice ale unui test — fie

Info 4.7

Considerații legate de fidelitate în selectarea testelor

Pasul 1	Determinați potențialele surse de eroare care pot să intre în scorurile instrumentelor examinate
Pasul 2	Examinați datele disponibile cu privire la fidelitate ale acestor instrumente, inclusiv tipurile de eșantioane din care au fost obținute datele.
Pasul 3	Evaluați datele fidelității în lumina tuturor celorlalte atribute ale testelor în chestiune, cum ar fi datele normative și cele de validitate, restricții de costuri și timp și așa mai departe.
Pasul 4	Dacă toate celelalte aspecte rămân neschimbate, selectați testul care promite să producă cele mai fidele scoruri în raport cu scopul testării și populația disponibilă.

că este vorba despre fidelitate, validitate, date normative sau orice alte aspecte tehnice ale unui instrument — nu există reguli fixe care se aplică tuturor testelor sau tuturor utilizărilor testelor.

Evaluarea surselor potențiale de eroare a scorurilor obținute la test

Cea mai importantă precauție necesară spre a minimiza eroarea scorurilor testului este de a adera în mod strict la procedurile standardizate de administrare și scorare a testelor (vezi Capitolul 7). Dincolo de aceasta, utilizatorii testului trebuie să evalueze relevanța posibilă a fiecăreia dintre sursele de eroare enumerate în Info 4.2 în vederea alegerii unor instrumente disponibile și adecvate scopurilor pentru care ele urmează a fi utilizate. De exemplu:

- Dacă scorarea unui test presupune o judecată subiectivă, trebuie luată în considerare fidelitatea inter-evaluatori.
- Dacă un test urmează a fi utilizat pentru a evalua schimbările în timp, cum ar fi posibilele ameliorări în urma intervenției terapeutice, este esențială estimarea erorii determinate de eșantionarea timpului, precum și a posibilelor efecte ale exersării asupra scorurilor obținute la instrumentele luate în considerare.
- Dacă există posibilitatea să fie nevoie ca o persoană să fie retestată la o dată ulterioară pentru a confirma sau valida rezultatele anterioare, este dezirabilă disponibilitatea unei forme alternative a testului, cu o fidelitate înaltă a scorurilor estimată prin metoda formelor alternative decalate în timp.
- Dacă se caută un test omogen și consistent, vor trebui căutați coeficienți K-R 20 sau coeficient alpha înalți.

Evaluarea datelor fidelității

Coeficienții de fidelitate oferă utilizatorilor testului anumite informații cu privire la mărimea erorii care apare din diferite surse și este probabil să intre în scoruri. Totuși, în evaluarea datelor fidelității trebuie reținut faptul că aceste estimări sunt influențate de caracteristici ale eșantionului pentru care au fost calculate și pot fi sau nu pot fi generalizate la alt grup de persoane testate. Printre altele, aceasta înseamnă că mici diferențe în mărimea coeficienților diferitelor teste nu au o semnificație la fel de mare ca alte considerații. Mai mult, în lumina varietății de factori care pot avea un efect asupra fidelității scorurilor obținute la test, există o recunoaștere din ce în ce mai mare a faptului că, atunci când raportează rezultate ale studiilor, cercetătorii trebuie să includă

în mod rutinier datele fidelității scorurilor pentru propriile lor eșantioane (vezi Baugh, 2003; Onwuegbuzie & Daniel, 2002).

Atunci când un test trebuie utilizat în evaluarea individuală, opusă cercetărilor ce presupun date de grup, importanța examinării critice a informațiilor publicate despre fidelitatea scorurilor înaintea selectării unui test nu poate fi suficient subliniată. Pe lângă evaluarea eșantioanelor din care au fost obținute datele cu privire la mărime, reprezentativitate și variabilitate, utilizatorii testului trebuie să ia în considerare cu atenție dacă coeficienții disponibili sunt cei mai adecvați pentru tipul de instrument accesibil și pentru modul în care intenționează să utilizeze testul. Mai mult, dacă un test este alcătuit din subteste sau din alte părți ale căror scoruri trebuie interpretate independent sau combinat, estimări ale fidelității pentru scorurile fiecărei părți trebuie să fie disponibile alături de estimările pentru scorul total al testului.

Am putea spune că un coeficient de fidelitate poate fi descris ca o corelație a testului cu sine însuși. Deși nu în întregime precisă, această descriere amintește faptul că coeficienții de fidelitate se bazează pe date — cum ar fi două administrări ale aceluiași test, două versiuni ale aceluiași test, corelații inter-item și așa mai departe — care *trebuie să fie* înalt consistente. Estimări scăzute ale fidelității (sub 0,70) sugerează faptul că nu se poate avea mare încredere în scorurile obținute la test. Astfel, deși nu există un prag minim pentru care un coeficient de fidelitate este considerat adecvat oricărui scop, se înțelege că, dacă toate celelalte aspecte rămân neschimbate, cu cât este mai mare coeficientul de fidelitate, cu atât este mai bine. Cei mai mulți utilizatori de teste caută coeficienți care să fie mai mari sau egali cu 0,80.

Evaluarea datelor scorurilor fidelității în lumina altor atribute ale testului

Deciziile de selectare a unui test trebuie luate pe baza fiecărui caz în parte, luând în considerare toate caracteristicile instrumentelor disponibile și a constructelor pe care își propun să le evalueze, ca și cerințele situației specifice în care vor fi utilizate scorurile testelor. Deși fidelitatea testelor este fundamentală, în niciun caz nu este singurul element luat în considerare în selectarea testelor. Pe lângă fidelitate, datele validității (Capitolul 5) și disponibilitatea informațiilor normative și a celor despre criteriu necesare interpretării scorurilor obținute la test (Capitolul 3) au o importanță deosebită. Deși considerațiile practice — cum ar fi costurile, ușurința administrării, restricțiile de timp și altele –joacă în mod necesar un rol în selectarea testului, în măsura în care este probabil ca folosirea unui test să aibă un impact semnificativ asupra persoanelor testate, aceste considerații nu ar trebui să fie factori determinanți în alegerea testului.

Evaluarea erorilor provenite din surse multiple

Cele mai multe scoruri obținute la test sunt susceptibile la erori de măsură care apar din mai multe surse. În teoria clasică a testării, această posibilitate realistă este ameliorată prin (a) metode care estimează influența combinată a două surse, cum ar fi fidelitatea prin forme alternative decalate în timp, care estimează atât eroarea determinată de eșantionarea timpului, cât și eroarea determinată de eșantionarea conținutului; sau (b) însumarea gradelor de variantă a erorii estimate prin toți coeficienții de fidelitate adecvați pentru a ajunge la o estimare a variantei totale a erorii. Ambele strategii se sprijină pe faptul că putem interpreta coeficienții de

fidelitate drept estimări ale proporției din dispersia scorurilor care poate fi atribuită unor surse diferite (vezi formula [4.4]). De exemplu, dacă coeficientul de fidelitate prin forme alternative decalate în timp al unui test este 0,75, atunci 75% din dispersia (varianța) scorurilor poate fi interpretată ca o dispersie reală și 25% (1-0,75 = 0,25) din dispersie poate fi atribuită influenței combinate a erorii de eșantionare a timpului și erorii de eșantionare a conținutului. Info 4.8 descrie o astfel de analiză a surselor de varianță a erorii pentru scorurile obținute la subtestul Vocabular din WAIS — III (Psychological Corporation, 1997).

Teoria generalizabilității

O abordare alternativă a fidelității care încearcă să fie mai cuprinzătoare decât cea pe care am discutat-o este cea numită *teoria generalizabilității*, sau, pur și simplu, *teoria G* (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). Teoria generalizabilității este o extindere a teoriei clasice a testării care utilizează metode de analiză dispersională (a varianței) (ANOVA) pentru a evalua simultan efectele combinate ale surselor multiple de varianță a erorii scorurilor testului. Un avantaj distinct al teoriei G — în comparație cu metodele de combinare a estimărilor fidelității ilustrate în Info 4.8 — este că permite evaluarea interacțiunilor efectelor ce provin din diferite surse de eroare. Prin urmare, este o tehnică de identificare a componentei varianței erorilor care poate intra în scoruri. Pe de altă parte, pentru a aplica modelul experimental cerut de teoria G, este necesar să obținem observații multiple pentru același grup de indivizi la toate variabilele independente care ar putea contribui la varianța erorilor în cazul unui test dat (scoruri în mai multe ocazii, între evaluatori, între diferite forme alternative etc). Cu

toate acestea, în general, atunci când este realizabil, rezultatele oferă o estimare mai bună a fidelității scorurilor decât abordările descrise anterior. În ciuda faptului că teoria G a fost introdusă inițial la începutul anilor '60, relativ puțini autori de teste au aplicat-o în construirea noilor instrumente. Totuși, odată cu răspândirea mai largă a acestei tehnici, este probabil ca popularitatea ei să crească. Cititorii care doresc să se familiarizeze cu tehnicile de bază ale teoriei G pot consulta o introducere concisă oferită de Thompson (2003a), care include un exemplu simplu de calcul. O abordare mult mai amplă și detaliată a cadrului conceptual și a aspectelor statistice ale teoriei generalizabilității poate fi găsită în volumul scris de Robert Brennan pe această temă (2001).

Abordarea fidelității în teoria răspunsului la item

Metode mai sofisticate de estimare a fidelității sunt disponibile prin teoria răspunsului la item (TRI) (introdusă în Capitolul 3 și prezentată în continuare în Capitolul 6). O explicație completă a aspectelor tehnice ale modelelor TRI depășește scopul acestui manual, dar avantajele pe care aceste modele le oferă, mai ales pentru testarea pe scară largă și testarea computerizată adaptativă, au stimulat rapid dezvoltarea și aplicarea lor în ultimele decenii. Prin intermediul metodelor TRI, fidelitatea și eroarea de măsură sunt abordate din punctul de vedere al funcției de informație a itemilor individuali ai testului, în contrast cu testul ca întreg. Deoarece nivelul de dificultate și puterea de discriminare a itemilor individuali — pentru trăsătura evaluată de test — pot fi calibrate cu mai mare atenție prin metode TRI, informația pe care răspunsul la test al fiecărui individ o oferă este mai precisă și, ca

Info 4.8

Analiza surselor multiple de varianță a erorii pentru scorurile unui singur test

Subtestul Vocabular al scalei de inteligență pentru adulți Wechsler — ediția a treia (WAIS-III) constă într-o serie de cuvinte din ce în ce mai dificile care sunt citite subiectului de către examinator și, simultan, sunt prezentate vizual în caietul testului. Definițiile subiectului, de obicei prezentate oral, sunt înregistrate verbatim și scorate imediat de către examinator, utilizând o scală de 2, 1 sau 0 puncte, în funcție de calitatea răspunsului dat de către cel examinat. În scorarea răspunsurilor, examinatorii sunt ghidați de o mare familiarizare cu eșantioanele de răspuns oferite în manual pentru fiecare dintre cuvinte — la cele trei niveluri ale scorului — și de definițiile din dicționar ale fiecărui cuvânt (Psychological Corporation, 1997).

Scorul total pentru subtestul Vocabular este suma punctelor câștigate de examinat la toți itemii rezolvați (cuvinte). Un scor de această natură este susceptibil la erori determinate de eșantionarea timpului și a conținutului și la posibilitatea existenței diferențelor inter-evaluatori. Estimările medii ale fidelității oferite în manualul WAIS-III (care pot fi consultate de cei care doresc mai multe informații) pentru subtestul vocabular sunt următoarele:

Sursa de eroare/ tipul de fidelitate	Coeficientul mediu	Proporția și procentul din varianța erorii
Eșantionarea timpului/ stabilitate (test-retest)	0,91	1-0,91 = 0,09 (9%)
Eșantionarea conținutului/ consistența internă	0,93	1 — 0,93 = 0,07 (7%)
Diferențe inter-evaluatori/ inter-evaluatori	0,95	1 — 0,95 = 0,05 (5%)
Totalul varianței erorii măsurate		0,09+0,07+0,05 = 0,21 (21%)
Varianța reală estimată		1-0,21 = 0,79 (79%)

Din calculele anterioare ar trebui să fie evident că, în cazul unui test de acest tip, exemplificat prin subtestul Vocabular al WAIS-III, dacă ne bazăm pe o estimare a fidelității pe baza unei *surse unice*, putem avea impresia greșită a unui posibil grad de eroare a scorurilor sale. Mai mult, acest exemplu evidențiază faptul că, pentru ca scorurile care sunt supuse unor surse de eroare multiple să prezinte suficientă încredere, estimările fidelității pentru fiecare sursă unică trebuie să fie destul de înalte, în intervalul 0,90 sau mai mare. Vezi Exemplul I: aplicarea SEM pentru o aplicație specifică a acestei analize a surselor multiple de eroare și a efectelor sale asupra fidelității unui scor din subtestul Vocabular din WAIS — III.

urmare, mai fidelă. În tipul de testare computerizată adaptativă pe care acestea o permit, selectarea celor mai adecvați itemi care urmează să fie prezentați subiecților este determinată de răspunsurile lor anterioare. Utilizând metodologia TRI și a testării adaptative, poate fi obținută o fidelitate adecvată cu o eroare de măsură minimă în cazul testelor mai scurte decât testele tradiționale (care oferă același conținut fix pentru toți subiecții), dat fiind că este disponibilă o bază de itemi suficient de extinsă și cuprinzătoare. Acesta este doar unul dintre numeroasele moduri fundamentale în care versiunea de măsurare bazată pe model cunoscută drept TRI diferă de regulile și principiile teoriei clasice a testării (vezi Embretson & Reise, 2000, p. 13–39).

Considerații legate de fidelitate în interpretarea unui test

Din momentul în care un test a fost ales, administrat și scorat, datele fidelității sunt aplicate în procesul de interpretare a testului

în două scopuri distincte, dar legate între ele. Primul este de a recunoaște și cuantifica intervalul de eroare al scorurilor obținute la test. Al doilea scop este de a evalua semnificația statistică a diferenței dintre scorurile obținute care ajută la determinarea importanței acestor diferențe în termenii a ceea ce reprezintă scorurile.

Cuantificarea erorilor din scorurile obținute la test: Eroarea standard de măsură

În interpretarea oricărui scor al unui test — sau a mediei scorurilor — datele de fidelitate sunt utilizate pentru a determina limita superioară și inferioară a intervalului în care este probabil să se situeze scorurile reale ale subiecților testați. Este calculat un interval de încredere pentru un scor obținut pe baza fidelității estimate a scorurilor obținute la testul în cauză. Mărimea intervalului depinde de nivelul de probabilitate ales.

Exemplul 1: Aplicarea SEM

Fidelitatea estimată a scorurilor la subtestul Vocabular din WAIS-III descrisă în Info 4.8 — după extragerea varianței estimate a erorii din trei surse relevante — este 0,79. Ca în cazul scorurilor obținute la toate subtestele Wechsler, Vocabularul are scoruri pe scale care se întind între 1 și 19, cu $M = 10$ și $SD = 3$. Pentru a ilustra aplicația fundamentală a datelor fidelității, să presupunem că o persoană numită Maria a obținut scorul 15 la subtestul Vocabular Wechsler.

Pasul 1. Pentru a obține un interval de încredere pentru scorul 15 obținut de Maria (X_o), avem nevoie de *eroarea standard de măsură* pentru Vocabular. *SEM* este o mărime statistică ce reprezintă

abaterea standard a distribuției ipotetice pe care am fi obținut-o dacă Mariei i s-ar fi aplicat acest test de un număr infinit de ori. Așa cum s-a menționat mai sus în acest capitol, media unei astfel de distribuții ipotetice ar fi *scorul real* al Mariei la subtestul Vocabular. Investigarea formulei prezentate în Info 4.9 arată că *SEM*

Info 4.9

Formula Erorii standard de măsură (SEM)

$$SEM = SD_t\sqrt{1 - r_{xx}}$$

unde
SD_t = abaterea standard a scorurilor obținute la test
r_{xx} = coeficientul de fidelitate

Formulele pentru eroarea standard a diferenței dintre două scoruri (SE_{dif})

Formula 1 SE_{dif}:

$$SE_{dif} = SD\sqrt{2 - r_{11} - r_{22}}$$

unde
SD = abaterea standard a testului 1 și a testului 2
r_{11} = estimarea fidelității pentru scorurile obținute la Testul 1
r_{22} = estimarea fidelității pentru scorurile obținute la Testul 2

Formula 2 SE_{dif}:

$$SE_{dif} = \sqrt{(SEM_1)^2 + (SEM_2)^2}$$

unde
SEM_1 = eroarea standard de măsură la Testul 1
SEM_2 = eroarea standard de măsură la Testul 2

Formula 1 SE_{dif} este utilizată dacă cele două scoruri la test care sunt comparate sunt exprimate pe aceeași scală, iar formula 2 SE_{dif} este utilizată dacă nu sunt exprimate în valori pe aceeași scală.

depinde de coeficientul de fidelitate al scorurilor testului și se exprimă în termeni de unități de abatere standard a scorurilor testului.

Din această cauză, mărimea *SEM* nu poate fi luată în sine ca un indicator al fidelității. Testele care au unități mari de abatere standard, cum ar fi SAT cu *SD* =100, vor avea *SEM* mult mai mari decât testele cu unități de abatere standard mici, cum ar fi subtestele scalei Wechsler cu *SD*= 3, deși coeficienții lor de fidelitate au mărime egală.

Pasul 2. Deoarece nu putem obține scoruri multiple la subtestul Vocabular pentru Maria, nici nu putem să facem o medie a lor pentru a obține o estimare a scorului real, trebuie să alegem un scor disponibil care poate fi plasat în centrul intervalului creat prin *SEM*. Aici apar două posibilități: (a) fie putem utiliza scorul obținut X_o ca un estimat al scorului real al Mariei, fie (b) putem estima scorul real (T') cu următoarea formulă, conform lui Dudek (1979):

$$T' = r_{xx}(X_o - M) + M \qquad (4.7)$$

unde

T' = scorul real estimat al individului
r_{xx} = fidelitatea estimată a scorurilor testului
X_o = scorul obținut de individ
M = media distribuției scorurilor obținute la test

În cazul Mariei, deoarece $X_o = 15$, $r_{xx} = 0{,}79$ și $M = 10$ pentru subtestul Vocabular, ca pentru toate scorurile subtestului Wechsler, scorul ei real estimat este 14 ($T' = (0{,}79)\ (15 - 10) + 10 = 13{,}95$ sau 14). Observați că, deoarece scorul ei este situat deasupra mediei, *scorul real estimat este mai mic decât scorul obținut*. Dimpotrivă, un

scor 5 obținut la același subtest — care se abate de la medie într-o măsură egală cu scorul Mariei, dar în direcția opusă — va duce la o estimare a scorului real de 6, care *mai mare* decât X_o (dacă $X_o = 5$, $T' = (0{,}79)\ (5 - 10) + 10 = 6{,}05$ sau 6). Motivul pentru această diferență între estimările scorurilor reale ale scorurilor obținute situate deasupra mediei și ale celor care se situează sub medie este că procedura de estimare ține cont de efectul regresiei către medie. Din același motiv, dacă $X_o = M$, cea mai bună estimare a scorului real va fi chiar media.

Pasul 3 . Dacă este sau nu necesar să se calculeze T' pentru a crea un interval de încredere depinde de măsura în care un scor obținut deviază de la medie. Dacă un scor obținut este mai apropiat de medie, scorul real estimat nu va diferi cu mult de ea; pe de altă parte, pe măsură ce scorurile obținute sunt mai apropiate de extreme, calcularea estimărilor scorurilor reale care sunt mai apropiate de medie este din ce în ce mai recomandată. Oricum ar fi, Pasul 3 presupune calcularea *SEM*. Utilizând formula din Info 4.9, descoperim că $SEM = \sqrt{1 - 0{,}79} = 1{,}37$. Deoarece această *SEM*, ca și celelalte erori standard descrise în Capitolul 2, reprezintă abaterea standard a distribuției ipotetice a scorurilor presupusă a fi normală, o putem interpreta în termenii frecvențelor curbei normale. Ne putem aminti, din Capitolul 3, că aproximativ 68% din aria de sub curba normală este cuprinsă în intervalul ±1*SD* de la medie, 95% în intervalul ± 1,96 *SD* și așa mai departe. Aplicând aceste procente scorului real estimat al Mariei (T') 14 și aplicând *SEM* obținută de 1,37, putem spune că (a) sunt șanse de 60/100 sau $p = 0{,}32$ ca scorul real al Mariei să se afle în intervalul 14 ± 1,37, care este între 13 și 15; și că (b) sunt 95/100 șanse, sau $p = 0{,}05$, ca scorul real să se afle în intervalul 14 ± (1,37) (1,96), adică între 11 și 17.

Scopurile evaluării presupun deseori comparații (a) între două sau mai multe scoruri obținute de același individ la diferite părți ale unei baterii de teste, ca în situația în care sunt comparate nivelele de performanță în diferite domenii, sau (b) între scorurile obținute de două sau mai multe persoane la același test, în scopul evaluării meritelor sau caracteristicilor lor relative. În ambele cazuri, datele fidelității pot fi utilizate pentru a emite afirmații cu privire la probabilitatea ca diferențele obținute între scoruri — și ceea ce reprezintă scorurile — să se datoreze întâmplării. Mărimea statistică utilizată în acest scop este *eroarea standard a diferențelor dintre scoruri* sau SE_{dif}. Aceasta poate fi calculată utilizând una dintre cele două formule prezentate în Info 4.9, ținând cont de faptul că scorurile ce urmează a fi comparate sunt exprimate pe aceeași scală (Formula 1) sau nu (Formula 2). Indiferent care dintre cele două formule este utilizată, SE_{dif} va fi mai mare decât *SEM* a fiecăruia dintre scorurile comparate deoarece evaluarea diferențelor dintre scoruri trebuie să țină cont de eroarea prezentă în ambele scoruri.

Exemplul 2: Aplicarea SE_{dif}

Pentru a ilustra utilizarea erorii standard a diferențelor dintre scoruri, să presupunem că dorim să estimăm semnificația statistică a diferenței dintre scorurile obținute de Maria la două subteste ale WAIS-III: scorul 15 la subtestul Vocabular și scorul 10 la subtestul Informații generale. Subtestul Vocabular este descris în Info 4.8; subtestul Informații evaluează cunoștințele despre evenimente comune, obiecte, locuri și oameni.

Pasul 1. Deoarece vrem să estimăm semnificația diferenței dintre două scoruri obținute, primul pas este de a calcula această diferență. În acest caz, 15 — 10 = 5. Există o diferență de 5 puncte între scorul Mariei la subtestul Vocabular și scorul ei la subtestul Informații generale. Știind asta, putem trece la a evalua dacă diferența obținută este semnificativă din punct de vedere statistic (adică este improbabil să apară întâmplător).

Pasul 2. Trebuie să calculăm eroarea standard a diferenței dintre scoruri la subtestele Vocabular și Informații generale din WAIS-III. Din moment ce cele două subteste sunt exprimate pe aceeași scală de scoruri ($M = 10$ și $SD = 3$), putem utiliza Formula 1 din Info 4.9. Această formulă ne cere să cunoaștem coeficienții de fidelitate pentru scorurile subtestelor. Coeficientul combinat pentru Vocabular este 0,79, estimat în Info 4.8. Pentru subtestul Informații generale, coeficientul estimat este 0,85, pe baza coeficienților combinați de consistență internă și stabilitate de 0,91 și, respectiv, 0,94, prezenți în *WAIS-III/WMS-III Technical Manual* (Psychological Corporation, 1997). Astfel, eroarea standard a diferenței dintre scorurile la subtestele Vocabular și Informații generale este

$$SE_{dif} = 3\sqrt{2 - 0{,}79 - 0{,}85} = 1{,}80$$

Pasul 3. Pentru a determina semnificația statistică a diferenței de 5 puncte între scorurile obținute, împărțim această diferență la SE_{dif} și obținem o valoare critică de 5/1,80 = 2,78.

Pasul 4. Consultând Tabelul ariilor situate sub curba normală din Anexa C pentru o valoare *z* de 2,78 descoperim că aria din porțiunea mai mică demarcată de acea valoare *z* este 0,0027. Deoarece nu există niciun motiv să presupunem că unul sau altul dintre scorurile obținute la subteste (Vocabular și Informații generale) ar fi mai mare decât celălalt, este adecvat un test bilateral

de semnificație pentru ipoteza de nul că nu există nicio diferență între scoruri. Prin urmare, înmulțim 0, 0027 cu 2 și obținem 0,0054 care indică faptul că probabilitatea ca scorurile Mariei la subtestele de Vocabular și Informații generale să difere cu 5 puncte datorită întâmplării este de 5,4 la 1000. Dat fiind acest nivel înalt de semnificație a diferenței, toate celelalte aspecte rămânând neschimbate, putem deduce cu siguranță că există *într-adevăr* o diferență: cunoștințele de vocabular ale Mariei, măsurate prin subtestul Vocabular al WAIS-III, depășesc cel mai probabil cunoștințele generale despre evenimente obișnuite, locuri și oameni măsurate prin subtestul Informații generale.

De ce este important să creăm intervale de încredere pentru scorurile obținute și pentru diferențele dintre scorurile obținute? Două motive fundamentale pot fi aduse ca răspuns la această întrebare. Primul este faptul că intervalele de încredere pentru scorurile obținute ne reamintesc faptul că scorurile testului nu sunt atât de precise pe cât ar părea să sugereze natura lor numerică. Astfel, ori de câte ori trebuie luate decizii importante cu ajutorul scorurilor obținute la teste, mai ales atunci când sunt utilizate scoruri bine delimitate, trebuie să se acorde o atenție deosebită erorii de măsură cuantificată prin *SEM*. Al doilea motiv, legat de primul, este că intervalele de încredere ne împiedică să atribuim o semnificație nepotrivită diferenței dintre scoruri care poate fi nesemnificativă dacă ținem cont de eroarea de măsură. Recunoscând importanța acestor fapte, multe manuale ale testelor includ tabele în care sunt prezentate erorile standard de măsură pentru scoruri și intervale numerice pentru fiecare scor care poate fi obținut la test, alături de nivelele de încredere pentru fiecare interval de scoruri. De exemplu, pentru un IQ total de 110 la WAIS-III, intervalul cu nivel de încredere de 90% este între 96 și 113 (Wechsler, 1997, p. 198).

Prezența acestei informații în manualul testului încurajează utilizatorii testului să aplice intervalele de încredere în interpretarea scorurilor fără a fi nevoie să le calculeze. Cu toate acestea, este sarcina utilizatorului să determine dacă aspectele publicate sunt aplicabile și au semnificație într-o anumită situație de utilizare a testului.

Tabelul 4.1 și Figura 4.1 ilustrează modul în care fidelitatea scorurilor și datele *SEM* pot fi utilizate în analiza a 4 — dintr-un total de 14 — scoruri ale subtestelor care pot fi obținute din WAIS-III. Pe lângă scorurile la subtestele Vocabular și Informații generale obținute de Maria, întâlnite deja în exemplele anterioare, două alte scoruri fictive ale subtestelor WAIS-III, și anume Raționament aritmetic și Memorare de cifre, au fost adăugate profilului ei. Aceste subteste au fost selectate pentru că aptitudinile primare pe care le evaluează — aptitudinea cantitativă și memoria auditivă de scurtă durată — sunt suficient de diferite pentru a face o comparație interesantă și rezonabilă între ele și alte două scoruri. Benzile de eroare calculate pentru un nivel de încredere de 90% pe baza *SEM* pentru respectivele subteste sunt prezentate în Tabelul 4.1. La rândul său, *SEM* a fost calculată pe baza *tuturor* elementelor relevante pentru fidelitate prezentate în *WAIS-III/WMS-III Technical Manual* pentru fiecare dintre cele patru subteste (Psychological Corporation, 1997). Această practică mai riguroasă este în contradicție cu utilizarea *SEM* (mai mici) bazate doar pe o estimare a fidelității, oferită în tabelele manualului testului pentru IQ și scorurile indicatorilor WAIS-III (vezi Wechsler, 1997, p. 195–202). Figura 4.1 prezintă datele din Tabelul 4.1 în formă grafică. Investigarea acestei figuri arată rapid faptul că, odată ce *SEM* este luată în calcul, intervalele probabile ale scorurilor Mariei se suprapun considerabil, ceea ce înseamnă că

Tabelul 4.1 Profilul scorurilor obținute de Maria la patru subteste WAIS-III cu SEM și benzi de eroare la un nivel de încredere de 90%

Subtest WAIS-III	Scorul obținut de Maria (X_o)	Coeficientul estimat de fidelitate[a]	*SEM*[b](1,64)[c] = banda de eroare	X_o+banda de eroare la nivel de încredere de 90%
Vocabular	15	0,79	1,37(1,64)= 2,25	15 ± 2,25= de la 12,75 la17,25
Raționament aritmetic	12	0,74	1,53(1,64)= 2,51	12 ± 2,51 = de la 9,49 la 14,51
Memoria cifrelor	17	0,83	1,24(1,64)= 2,03	17 ± 2,03 = de la 14,97 la 19,01
Informații generale	10	0,85	1,16(1,64)= 1,90	10 ± 1,90 = de la 8,10 la 11,90

Notă: WAIS-III = Scala de inteligență Wechsler pentru adulți — Ediția a treia; *SEM* = eroarea standard de măsură; X_o = scor observat; *SD* = abatere standard.

[a] Fidelitatea estimată după extragerea tuturor estimărilor pertinente ale variației erorii așa cum se vede în exemplul din Info 4.8.

[b] $SEM = SD\sqrt{1 - r_{xx}}$

[c] 1,64 este valoarea *z* pentru $p = 0,10$ (nivel de încredere 90%)

anumite diferențe între scorurile obținute se pot datora erorii de măsură. În același timp, acest tip de analiză de profil permite utilizatorului testului să exploreze *ipotezele* legate de semnificația posibilă a diferențelor performanțelor Mariei la abilitățile surprinse

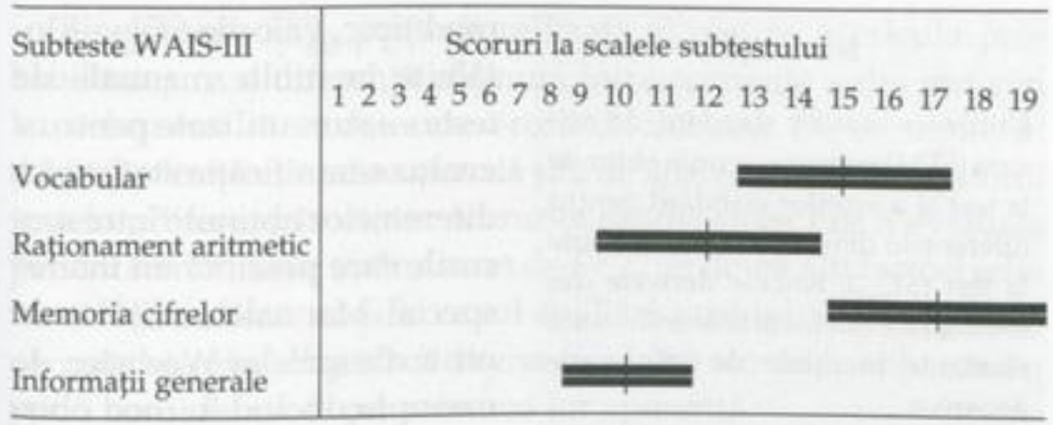

Figura 4.1 Profilul grafic al scorurilor obținute de Maria la subteste, ilustrând utilizarea benzilor de eroare din Tabelul 4.1.

de subtestele ale căror benzi de eroare nu se suprapun (i.e. Vocabular și Informații generale, Raționament aritmetic și Memoria cifrelor și Memoria cifrelor și Informații generale). De exemplu, se pare că Maria ar putea prezenta un punct forte relativ în capacitatea memoriei de scurtă durată și a vocabularului, în timp ce nivelul de informații generale poate fi un punct relativ slab. Bineînțeles, o concluzie bazată pe astfel de diferențe trebuie supusă confirmării sau revizuirii în lumina unor date suplimentare. Cu toate acestea, atunci când evaluarea psihologică cere o determinare a posibilelor puncte puternice și puncte slabe ale individului în aria intelectuală sau în alte domenii — de exemplu, interes vocațional — acest tip de analiză exploratorie, deși nu definitivă, poate fi destul de utilă.

Eroarea standard a diferențelor dintre scoruri (SE_{dif}) discutată anterior servește unui scop similar analizei de profil prezentată în Figura 4.1. Aceasta oferă date cu privire la discrepanțele dintre scoruri care pot avea o semnificație practică sau psihologică. În

Nu uitați!

Utilizarea erorilor standard de măsură (SEM) pentru scorurile obținute la test și a erorilor standard pentru diferențele dintre scorurile obținute la test (SE_{dif}), ambele derivate din estimări ale fidelității scorurilor, sunt elemente esențiale de informație deoarece

1. *SEM* oferă intervale de încredere pentru scorurile obținute la test care avertizează utilizatorii testului cu privire la faptul că scorurile sunt supuse unor fluctuații datorate erorii de măsură și
2. Intervalele de încredere obținute prin utilizarea statisticilor SE_{dif} împiedică supraevaluarea diferențelor dintre scoruri care pot fi nesemnificative în lumina erorii de măsură.

mod tipic, valorile SE_{dif} — întâlnite în multe manuale de teste — sunt utilizate pentru a evalua semnificația statistică a diferențelor obținute între scorurile care prezintă un interes special. Manualele ediților recente ale scalelor Wechsler, de exemplu, includ în mod obișnuit tabele care prezintă diferențele punctuale de valori IQ (IQ verbal versus IQ de performanță) și indexul scorurilor necesare pentru o semnificație statistică la intervalele de încredere de 90% și 95%. Recunoscând faptul că o diferență semnificativă *din punct de vedere statistic* poate să nu fie neapărat semnificativă *din punct de vedere psihologic*, autorii acestor manuale ale testului — și cei ai altor teste — oferă și tabele care arată frecvențele cu care diferențele între scoruri de diferite amplitudini au fost găsite în eșantionul de standardizare a testului în cauză. Această informație aduce în atenție întrebarea cât de obișnuită sau cât de rară este o anumită diferență dată în eșantionul normativ. Importanța sa apare din presupunerea (nu întotdeauna justificată, dar care merită investigată) că dacă diferențele de un anumit grad apar frecvent, ele au probabil o semnificație interpretativă mai mică decât cele care apar rar.

Bineînțeles, SE_{dif} poate fi utilizată și pentru a calcula probabilitatea ca o diferență obținută între scorurile a doi indivizi la același test să se datoreze erorii de măsură. De exemplu, în deciziile de selecție educațională sau de angajare luate cu ajutorul testelor, diferențele dintre scorurile candidaților pot fi evaluate pentru semnificație ținând cont de SE_{dif} pe lângă alți factori relevanți. Încă o dată, ca și în cazul analizei profilului, acest proces necesită atenție la faptul că diferențele între scorurile obținute nu pot fi considerate după valoarea lor aparentă.

Relația dintre fidelitate și validitate

Din perspectivă psihometrică, dovezile fidelității scorurilor sunt considerate a fi o condiție necesară, dar insuficientă a validității (vezi Sawilowsky, 2003). De fapt, așa cum vom vedea în capitolul următor, cele două concepte au o legătură intrinsecă prin faptul că fidelitatea scorurilor în sine poate fi văzută ca o dovadă minimă că a fost atinsă o măsură validă a unui eșantion de comportament.

În general, profesioniștii din evaluarea psihologică consideră că dovada fidelității scorurilor nu reprezintă o bază suficientă pe care să facă inferențe valide despre semnificația scorurilor. Cu toate acestea, există un anumit dezacord cu privire la măsura în care demonstrarea fidelității este considerată *esențială* pentru o evaluare validă a tuturor tipurilor de eșantioane de comportament care pot fi incluse în teste. De exemplu, atunci când scorurile la test sunt obținute din eșantioane de comportament unice sau idiosincratice, ele pot să nu fie repetabile sau constante. Testele care cer un nivel de performanță optimă a individului, cum ar fi eșantioane de activitate profesională sau portofolii, pot duce la

rezultate valide și fidele în termenii corectitudinii și preciziei, dar nu în termenii consistenței sau stabilității (vezi Moss, 1994). În mod similar, instrumentele care sunt administrate individual, ca multe scale de inteligență sau tehnici proiective, sunt susceptibile într-un grad ridicat la influențele ce apar din calitatea raportului dintre examinator și examinat, ca și din alți factori motivaționali și situaționali. În contextul evaluării individuale, astfel de instrumente pot oferi imagini valide ale aspectelor alcătuirii psihice a unei persoane care nu ar putea fi reproduse cu un examinator diferit sau în circumstanțe diferite, chiar dacă procedurile standardizate sunt observate în mod rigid (Masling, 1960; McClelland, 1958; Smith, 1992).

Concluzie

Utilizarea testelor psihologice ar fi mult simplificată dacă coeficienții de fidelitate și *SEM* ar putea fi luate în considerare după valoarea lor aparentă în evaluarea scorurilor testului. Așa cum atestă acest capitol, totuși, fidelitatea scorurilor reprezintă o judecată relativă, bazată atât pe date psihometrice, cât și pe contextul în care este administrat testului. Vom vedea în Capitolul 5 că același lucru este adevărat și pentru datele validității scorurilor obținute la test. Astfel, deși disponibilitatea datelor psihometrice adecvate despre fidelitatea scorurilor este o cerință fundamentală pentru orice utilizare a scorurilor testului, contextul în care se desfășoară testarea psihologică este, de asemenea, un aspect fundamental al interpretării scorurilor obținute la test de indivizi sau grupuri. Pe măsură ce impactul posibil al deciziilor luate cu ajutorul scorurilor testului crește, ambii factori câștigă o importanță mai mare.

TESTEAZĂ-TE SINGUR

1. **Un scor real este**
   a. o entitate ipotetică
   b. o entitate reală
   c. egal cu scorul observat
   d. egal cu scorul observat plus eroarea

2. **Dacă fidelitatea unui test este bine stabilită, utilizatorul testului poate presupune că scorurile obținute la acel test vor fi fidele?**
   Adevărat sau fals?

3. **Care dintre următoarele surse de eroare a scorurilor testului *nu* este evaluată prin estimările tradiționale ale fidelității?**
   a) diferențele inter-evaluatori
   b) eșantionarea timpului
   c) eșantionarea conținutului
   d) deviațiile de la procedurile standardizate

4. **Coeficienții de fidelitate_____ sunt utilizați pentru a estima eroarea determinată de eșantionarea timpului în scorurile testului.**
   a) test-retest
   b) forme alternative
   c) inter-evaluatori
   d) prin înjumătățire

5. **Care dintre următoarele tipuri de coeficienți de fidelitate duce la o estimare combinată a erorii apărute din două surse diferite?**
   a) Inter-evaluatori
   b) Test-retest
   c) Forme alternative
   d) Forme alternative decalate în timp

6. **Dacă toate celelalte aspecte rămân neschimbate, scorurile obținute la teste mai lungi sunt _____ cele obținute la teste comparabile care sunt mai scurte.**
   a) mai puțin fidele decât
   b) mai fidele decât
   c) la fel de fidele ca și

7. **Este mai probabil ca mărimea coeficientului de fidelitate să fie influențată de _____ decât de _____ eșantionului din care este calculat.**
   a) mărime/eterogenitate
   b) eterogenitate/mărime

8. **Un avantaj distinct al teoriei generalizabilității față de abordările tradiționale ale fidelității scorurilor este că teoria generalizabilității**
   a) necesită un număr mai mic de observații.
   b) duce la componente mai mici de eroare.
   c) permite evaluarea interacțiunii efectelor.
   d) utilizează metode statistice mai puțin complicate.

9. **Să presupunem că un student obține un scor de 110 la un test cu $M = 100$, $SD = 20$ și o fidelitate estimată de 0,96. Sunt 68 din 100 șanse ca scorul real al studentului să intre în intervalul**
   a) 100 și 110
   b) 102 și 112
   c) 106 și 114
   d) 110 și 120

10. **Eroarea standard de măsură a Testului A este 5 și eroarea standard de măsură a Testului B este 8. Eroarea standard a diferenței pentru compararea scorurilor la cele două teste va fi**
    a) mai mică de 8.
    b) mai mică decât 5.
    c) între 5 și 8.
    d) mai mare de 8.

Răspunsuri: 1. a; 2. b; 3. d; 4. a; 5. d; 6. b; 7. b; 8. c; 9. c; 10. d.

Aspecte esențiale ale validității

Testele psihologice există pentru a ne ajuta să facem inferențe cu privire la oameni și comportamentul lor. Validitatea — care este, de departe, aspectul fundamental legat de scorurile obținute la test și de utilizările lor — se sprijină pe dovezile pe care le putem aduce pentru a susține orice inferență făcută pe baza rezultatelor la test. Primatul considerațiilor legate de validitate este recunoscut în volumul actual al *Testing Standards* prin plasarea acestei teme în primul capitol, care definește *validitatea* ca fiind „gradul în care toate dovezile acumulate susțin interpretarea pe care intenționăm să o dăm scorurilor obținute la teste, conform scopului propus" (AERA, APA, NCME, 1999, p. 11). Implicit, în această definiție există trei idei interrelaționate care reflectă perspectivele actuale din testarea psihologică asupra acestui concept central și multifațetat.

1. Validitatea scorurilor obținute la test derivă din toate datele acumulate pentru a susține interpretarea și utilizarea lor. Astfel, validitatea este întotdeauna o problemă de grad, și nu o determinare de tipul totul sau nimic. *Validarea* — procesul

prin care sunt acumulate dovezi ale validității — începe cu declararea explicită de către autorul testului a cadrului conceptual și a argumentelor care susțin testul, dar, prin natura sa, este deschisă, deoarece include toate informațiile care se adaugă modului în care înțelegem rezultatele obținute la test.

2. Pe măsură ce înțelegerea teoretică și dovezile empirice ale interpretărilor scorurilor testului se acumulează, validitatea inferențelor (i.e. ipotezelor) făcute pe baza scorurilor testului în diferite scopuri poate fi amplificată sau diminuată. Un corolar al acestei noțiuni, afirmat explicit în *Testing Standards* (AERA, APA, NCME, 1999), este faptul că „validarea este responsabilitatea comună a celui care construiește testul [care oferă dovezi și argumentează modul în care intenționează să utilizeze testul] și a utilizatorului testului [care evaluează dovezile disponibile în contextul în care testul urmează a fi utilizat]" (p.11).
3. Deoarece scorurile obținute la test pot fi aplicate în multe scopuri diferite, bazele care susțin interpretarea scorurilor pot fi obținute prin diferite metode. Contribuții la dovezile validității scorurilor unui test pot fi aduse prin orice cercetare sistematică care susține sau adaugă ceva semnificației lor, indiferent de persoana care conduce această cercetare sau momentul în care are loc cercetarea. Atât timp cât există dovezi științifice convingătoare pentru o utilizare propusă a scorurilor testului, utilizatorii calificați pot utiliza scorurile în scopuri proprii, indiferent dacă acestea au fost sau nu prevăzute de autorii testului. Această afirmație ajută la explicarea naturii multifațetate a cercetărilor de validare, ca și a rezultatelor lor deseori redundante și uneori contra-

Nu uitați!

Poate că niciun alt teoretician nu a avut o influență mai mare decât Samuel Messick în remodelarea conceptului de validitate. Conform lui Messick (1989, p.13) „validitatea este o judecată evaluativă integrată a gradului în care dovezile empirice și argumentele teoretice susțin *adecvarea* și *recomandabilitatea inferențelor* și *acțiunilor* bazate pe scorurile obținute la test sau pe alte moduri de evaluare."

dictorii. De asemenea, explică longevitatea anumitor instrumente, cum ar fi MMPI și scalele Wechsler, pentru care s-a acumulat o literatură vastă — cuprinzând numeroase aplicații în diferite contexte — în decenii de cercetări fundamentale și aplicate.

Cititorul atent a putut înțelege până în acest punct că validitatea, ca și fidelitatea, nu este o calitate ce caracterizează testele în mod abstract sau doar anumite teste sau date specifice ale testului. Validitatea este mai mult o problemă de *judecată* care ține de utilizarea scorurilor obținute la test într-un anumit scop și într-un context dat. Astfel, procesul de validare este similar cu testarea unei ipoteze: subsumează noțiunile de semnificație a scorurilor obținute la test și de fidelitate a scorurilor testului discutate în cele două capitole anterioare, ca și modurile în care pot fi justificate aplicații ale datelor testului în cercetarea și practica psihologică, subiect de care se ocupă capitolul de față. Info 5.1 enumeră câteva dintre cele mai semnificative contribuții la tema validității din anii '50 până în anii '90.

Perspective istorice asupra validității

Apariția testării psihologice moderne a avut loc în perioada în care psihologia devenea o disciplină științifică. Ambele domenii își au începutul în ultima parte a secolului al XIX-lea și începutul secolului XX. Ca rezultat al acestei coincidențe istorice, modul în care înțelegem natura, funcțiile și metodologia testelor și instrumentelor psihologice a evoluat în ultimul secol în tandem cu dezvoltarea și sofisticarea din ce în ce mai mare a psihologiei ca știință.

La început, psihologia științifică era preocupată în primul rând cu stabilirea legilor psihofizice, prin investigarea experimentală a relației funcționale dintre stimulii fizici și răspunsurile senzoriale și perceptive pe care le trezesc oamenilor. Psihologia teoretică consta în principal din speculații de natură filosofică, până în primul sfert al secolului XX. Niciuna dintre aceste afirmații nu presupune lipsa de valoare a contribuțiilor aduse de pionierii psihologiei (vezi Boring, 1950; James, 1890). Totuși, pe acest fundal, primele teste psihologice au ajuns să fie privite, oarecum naiv, drept instrumente științifice care măsurau un catalog în permanentă extindere de aptitudini mentale și trăsături ale personalității cam în același mod în care psihofizicienii măsurau răspunsurile auditive, vizuale și alte răspunsuri senzoriale și perceptive la stimuli cum ar fi sunete, lumină și culori de diferite tipuri și intensități. Mai mult, așa cum am văzut în Capitolul 1, succesul pe care Stanford-Binet și Army Alpha l-au avut, susținând luarea unor decizii practice cu privire la indivizi în cadre educaționale și ocupaționale, a condus la o rapidă proliferare a testelor în primele două decade ale secolului XX. Gama largă de aplicații pentru care au fost utilizate aceste instrumente a depășit cu mult argumentele teoretice și științifice disponibile în acel moment. Pe

Info 5.1

Referințe de bază despre validitate

Samuel Messick și-a formulat în cel mai explicit mod punctul de vedere asupra validității într-un capitol apărut în *Educational Measurement* (ediția treia, pp. 13–103), un volum notabil editat de Robert L. Linn și publicat în colaborare de American Council on Education și Macmillan în 1989. Capitolul despre validitate scris de Messick și celelalte lucrări cu această temă (ex.: Messick, 1988, 1995), au influențat direct modul în care validitatea este tratată în versiunea actuală a *Testing Standards* (AERA, APA, NCME, 1999). Printre alte contribuții cheie larg recunoscute ca modelând evoluția conceptelor teoretice ale validității se numără:

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph Supplement]. *Psychological Reports, 3,* 635–694.
- Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.
- Cronbach, L. J. (1988). Five perspectives on validity argument. În H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

scurt, multe dintre primele teste psihologice au fost construite și utilizate fără a beneficia de teoria psihometrică, de principii etice și de ghiduri practice care au început să se acumuleze în deceniile următoare (von Mayrhauser, 1992).

Definiția clasică a validității

Recunoașterea acestei stări de lucruri în cadrul profesional a dus la primele încercări de a contura caracteristicile care ar putea face distincția între un test bun și unul prost. Astfel, prima definiție

a *validității* ca „măsura în care un test măsoară ceea ce și-a propus să măsoare" a fost formulată în 1921 de National Association of the Directors of Educational Research (T. B. Rogers, 1995, p. 25). A fost ratificată de mulți experți în testare — inclusiv de Anne Anastasi în toate edițiile manualului ei deosebit de important *Psychological Testing* (1954–1988) ca și în Anastasi și Urbina (1997, p. 8). Perspectiva conform căreia „validitatea testului privește *ceea ce* măsoară testul și *cât de bine* face acest lucru" (Anastasi & Urbina, p. 113) este privită de mulți drept aspectul central al problemei validității. În ciuda aparentei sale simplități, această abordare pune un număr de probleme, mai ales atunci când este privită din perspectiva ediției actuale a *Testing Standards* (AERA, APA, NCME, 1999) și a schimbărilor continue cu privire la definirea unora dintre constructele fundamentale din câmpul psihologiei.

Aspecte problematice ale perspectivei tradiționale asupra validității

Problemele ridicate de definiția clasică a validității pivotează în jurul unor premise neformulate, dar clare, conform cărora:

1. Validitatea este o proprietate a testelor, și nu a interpretării scorurilor obținute la un test;
2. Pentru a fi valide, scorurile testelor trebuie să măsoare direct un anumit construct urmărit; și
3. Validitatea scorurilor depinde, cel puțin într-o anumită măsură, de modul în care autorul testului înțelege constructul pe care intenționează să-l măsoare.

Deși aceste premise pot fi justificate în anumite cazuri, în mod clar ele nu se justifică în fiecare caz. Prima, de exemplu, poate fi

îndeplinită numai dacă datele validității susțin scopul declarat al testului *și* atât timp cât testul este utilizat în mod specific în acel scop și cu tipul de populație pentru care au fost acumulate date ale validității. A doua și a treia premisă se justifică doar pentru testele care măsoară comportamentul ce poate fi legat de constructe psihologice în moduri lipsite de echivoc, cum ar fi anumite funcții ale memoriei, viteza și precizia în performanța la diferite sarcini de procesare cognitivă sau extinderea cunoștințelor despre un domeniu de conținut bine definit. Ele nu pot fi menținute în mod necesar în cazul (a) testelor proiectate să evalueze constructe teoretice multidimensionale sau complexe în legătură cu care există încă largi dezbateri, cum ar fi inteligența sau conceptul de sine; (b) teste construite pe baza unei relații strict empirice — opusă celei teoretice sau logice — între scoruri și criteriul extern, cum ar fi versiunea inițială MMPI; sau (c) tehnici al căror scop este să dezvăluie aspecte ascunse sau inconștiente ale personalității, cum sunt tehnicile proiective. În cazul instrumentelor de acest tip, ceea ce se măsoară este comportamentul care poate fi legat mai mult sau mai puțin direct de constructe de un interes real, în primul rând printr-o rețea de dovezi corelaționale. Info 5.2 definește diferite semnificații ale cuvântului *construct* și poate ajuta la clarificarea distincțiilor pe care tocmai le-am făcut și a celor care vor apărea mai târziu pe parcursul acestui capitol.

Ideea că validitatea scorurilor testului depinde de gradul în care testul măsoară ceea ce și-a propus să măsoare conduce și la o anumită confuzie între consistența sau precizia instrumentelor (adică fidelitatea lor) și validitate. Așa cum am văzut în capitolul 4, dacă un test măsoară ceva *bine*, scorurile sale pot fi fidele (consistente, precise, de încredere), dar nu sunt în mod necesar

Info 5.2

Analiza constructelor

Deoarece termenul *construct* este utilizat atât de des în capitolul de față, în acest moment este necesară o clarificare a semnificației sale. În general vorbind, un *construct* este ceva inventat de mintea umană, care nu este observabil în mod direct. Constructele sunt abstractizări care se pot referi la concepte, idei, entități teoretice, ipoteze sau diferite invenții.
În psihologie, cuvântul *construct* se aplică conceptelor, cum ar fi trăsăturile, și relațiilor teoretice între concepte care sunt deduse din observații empirice consistente ale datelor comportamentale. Constructele psihologice diferă larg în termenii:

- ariei și complexității lor
- aplicabilității potențiale
- gradului de abstractizare necesar pentru a le deduce din datele disponibile.

Ca regulă, constructele cu arie restrânsă necesită o abstractizare redusă, dar au o gamă mai săracă de aplicații. Mai mult, deoarece este mai ușor să obții un acord consensual cu privire la constructe care sunt restrânse, simple și mai puțin abstracte, acestea sunt și mai ușor de evaluat decât constructele mai cuprinzătoare și multifațetate ce ar putea dobândi semnificații diferite în contexte, culturi și perioade istorice diverse.

Exemple:

- În timp ce *dexteritatea manuală* este un construct care poate fi legat imediat de date comportamentale specifice, *creativitatea* este un construct mult mai abstract. Ca urmare, când se pune problema evaluării acestor trăsături, este mult mai ușor să determini cine are o dexteritate manuală mai mare decât să stabilești cine este mai creativ.
- *Introversia* este un construct simplu și mai restrâns definit decât *capacitatea de conștientizare*. Deși ultima este potențial utilă în predicția unei game largi de comportamente, ea este și mai dificil de evaluat.

Sinonime: Termenii *construct* și *variabilă latentă* sunt deseori utilizați interșanjabil. O *variabilă latentă* este o caracteristică despre care se presupune că este subiacentă anumitor fenomene observabile, dar nu este observabilă și măsurabilă în mod direct. Toate trăsăturile psihice sunt variabile latente sau constructe, tot așa cum sunt etichetele aplicate unor factori care apar din cercetarea analitică factorială, cum ar fi comprehensiunea verbală sau nevroticismul.

valide în sensul contemporan, complet al termenului. Cu alte cuvinte, scorurile testului pot să fie relativ lipsite de erori de măsură, și totuși să nu fie foarte utile ca bază pentru realizarea inferențelor pe care trebuie să le facem.

Mai mult, presupunerea că scorurile unui test reflectă ceea ce autorul testului intenționează ca ele să reflecte a fost o sursă suplimentară de înțelegeri eronate. Una dintre ele privește titlurile testelor, ce nu ar trebui să fie niciodată luate în considerare după aparență — deși deseori se întâmplă astfel. Titlurile testelor variază de la cele care sunt destul de exacte și pot fi susținute empiric, la cele care reflectă doar intențiile (nerealizate) ale autorilor sau interesele de marketing ale editurilor care publică teste. A doua problemă, mai importantă, este legată de ideea că scorurile valide ale unui test reflectă scopul lor exprimat și poate conduce la definiții superficiale sau inexacte ale constructelor psihologice. Poate că cel mai faimos exemplu este definiția *inteligenței* dată în 1923 de E. G. Boring: „tot ceea ce măsoară un test de inteligență" (citat de Sternberg, 1986, p. 2). Ca rezultat al acestor înțelegeri greșite, domeniul testării psihologice a fost împovărat cu instrumente — care își propuneau să măsoare constructe

prost definite sau excentrice — ale căror promisiuni declarau în mare măsură mai mult decât puteau oferi, a căror utilizare în cercetarea psihologică a împiedicat sau a întârziat progresul disciplinei și a căror existență, prin asociere, a diminuat imaginea întregului domeniului. Principalele exemple ale acestui tip de problemă sunt primele măsurări ale masculinității — feminității (Constantinople, 1973; Lenney, 1991; Spence, 1993), deși există multe altele.

Poate că cea mai semnificativă consecință a definiției tradiționale a validității este că ea s-a atașat de teste și de ceea ce acestea își propun să măsoare, și nu de *scorurile* obținute la teste și de interpretările care s-ar putea baza în mod justificabil pe acestea. Prin urmare, orice dovadă etichetată drept *validitate a testului* a ajuns să fie văzută drept o dovadă că testul era valid și merita utilizat, indiferent de natura legăturii dintre datele scorurilor obținute la test și inferențele care urmau să fie făcute pornind de la acestea. În consecință, nenumărate studii publicate în literatura de specialitate au utilizat scoruri provenite de la un singur instrument pentru a clasifica participanții la cercetări în grupuri experimentale, mulți clinicieni s-au bazat exclusiv pe scorurile obținute la test pentru diagnoză și planificarea tratamentului și numeroase decizii din cadrul educațional și ocupațional s-au bazat pe scoruri distincte obținute la un singur test. Prea des astfel de alegeri s-au făcut fără a se lua în considerare adecvarea lor în contexte specifice sau fără o referință la surse suplimentare de date, fiind justificate pur și simplu prin faptul că testul în discuție era presupus a fi „o măsură validă a..." ceea ce susținea manualul său.

Un moment important în evoluția conceptului de validitate a fost publicarea *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954), primul din seria de standarde de testare care au fost redenumite, revizuite și aduse la zi în 1955, 1966, 1974, 1985 și, cel mai recent, în 1999 (AERA, APA, NCME). La fiecare revizuire, *Testing Standards* — discutat anterior în capitolul 1 — a încercat să promoveze practici competente pentru construirea și utilizarea testelor și să clarifice bazele evaluării calității testelor și metodelor de testare.

Technical Recommendations publicată în 1954 a introdus o clasificare a validității în patru categorii ce vor fi prezentate mai târziu în acest capitol: validitatea de conținut, predictivă, concurentă și validitatea de construct. Ulterior, *Standards* din 1974 a redus aceste categorii la trei, subsumând validitatea predictivă și validitatea concurentă în rubrica validitate în raport cu criteriul, și specificând mai departe că validitatea de conținut, de construct și cea în raport cu criteriul sunt *aspecte ale* validității, opuse *tipurilor de* validitate. În același an, *Standards* a introdus și ideea că validitatea „se referă la adecvarea inferențelor făcute pornind de la scorurile testului sau de la alte forme de evaluare" (APA, AERA, NCME, 1974, p. 25).

În ciuda specificărilor propuse de *Standards* din 1974 acum mai mult de un sfert de secol, împărțirea validității în trei tipuri (cunoscută drept *perspectiva tripartită* asupra validității) s-a consolidat. A supraviețuit până în prezent în multe manuale ale testelor și în recenzii de teste, ca și în multe cercetări asupra instrumentelor psihometrice. Cu toate acestea, revizuirile succesive ale *Standardelor* — mai ales cea actuală — au adăugat stipulații care fac din ce în ce mai clar faptul că oricare ar fi clasificarea utilizată pentru validitate, conceptele ar trebui atașate tipurilor de date care susțin interpretarea scorurilor testului, și nu testelor în sine. Reținând acest aspect, vom lua în considerare în continuare perspectiva prevalentă asupra validității ca un concept unitar și diferitele surse

de dovezi ce pot fi utilizate pentru a evalua interpretările posibile ale scorurilor unui test în scopuri specifice. Pentru mai multe informații despre evoluția validității și a conceptelor legate de aceasta, vezi Anastasi (1986), Angoff (1988) și Landy (1986).

Perspective actuale asupra validității

Începând din anii '70 și continuând până în prezent, în cadrul profesiei de testare psihologică s-a depus un efort concertat pentru a rafina și a revizui noțiunea de validitate și pentru a oferi o teorie unificatoare care să cuprindă numeroasele straturi de dovezi din care scorurile testelor își derivă semnificația și sensul. O temă constantă a acestui efort a fost integrarea tuturor formelor de dovezi ale validității ca aspecte ale validității de construct (Guion, 1991; Messick, 1980, 1988, 1989; Tenopyr, 1986). La rândul său, aceasta a dus la reexaminarea semnificației termenului *construct* — definit în termeni generali în Info 5.2 — așa cum este aplicat în mod specific în contextul validității în testarea și evaluarea psihologică (vezi Braun, Jackson, & Wiley, 2002; Embretson, 1983).

Funcția integrativă a constructelor în validarea testelor

În testarea psihologică, termenul *construct* a fost utilizat, deseori fără diferențiere, în două moduri alternative:

1. Pentru a denota *trăsăturile, procesele, bagajul de cunoștințe* sau *caracteristicile* a căror prezență și extindere dorim să o stabilim prin eșantioane de comportament specific colectate de teste. În acest sens al cuvântului, un construct este pur și simplu ceea ce autorul testului stabilește să măsoare — adică orice entitate ipotetică derivată din teoria psihologică, din cercetarea sau observarea comportamentului, cum ar fi anxietatea, asertivitatea, capacitatea de raționament logic, flexibilitatea și așa mai departe.
2. Pentru a denota *inferențele* ce pot fi făcute pe baza scorurilor testului. Atunci când este utilizat în acest mod, termenul *construct* se referă la o interpretare specifică a datelor unui test sau a oricăror alte date comportamentale — cum ar fi prezența depresiei clinice sau a unei probabilități mari de succes într-o anumită sarcină — care pot fi făcute pe baza unei rețele de relații teoretice și empirice prestabilite între scorurile testului și alte variabile.

Câțiva teoreticieni au încercat să explice modul în care aceste două sensuri se leagă de noțiunea de validitate a scorurilor testului. Una dintre primele formulări a fost clasificarea validității realizată de Cronbach (1949) în două categorii, și anume logică și empirică. Ulterior, într-un articol important pe care l-a scris împreună cu Meehl în 1955, Cronbach a sugerat utilizarea termenului *validitate de construct* pentru a desemna *rețeaua nomologică*, sau rețeaua de interrelații dintre elementele teoretice și elementele observabile care susțin un construct. În încercarea de a clarifica modul în care aceste două semnificații pot fi distinse în procesul de dezvoltare, construire și evaluare a testului, Embretson (1983) a propus o separare între două aspecte ale cercetărilor de validare de construct și anume, reprezentarea constructului și domeniul nomotetic. Conform lui Embretson (p. 180), cercetarea *reprezentării constructului* „se ocupă de identificarea mecanismelor teoretice subiacente performanței la o sarcină." Din perspectiva procesării

informației, scopul reprezentării constructului este *descompunerea sarcinii*. Procesul de descompunere a sarcinii poate fi aplicat unor diferite sarcini cognitive, inclusiv inferențe interpersonale sau judecăți sociale. Descompunerea presupune o examinare a răspunsurilor la test din punctul de vedere al proceselor, strategiilor și cunoștințelor implicate în performanță. *Domeniul nomotetic*, pe de altă parte, privește „rețeaua de relații ale unui test cu alte instrumente" (Embretson, p. 180); se referă la relațiile semnificative dintre scorurile testului și alte instrumente care măsoară aceleași trăsături — sau trăsături diferite — între scorurile la test și instrumentele criteriului și așa mai departe.

Embretson (1983) a descris caracteristicile suplimentare ale conceptelor de reprezentare a constructului și domeniul nomotetic ce ajută la clarificarea diferențelor între aceste două aspecte ale cercetărilor de validare a constructului. Două dintre punctele pe care le-a introdus privesc distincția dintre funcțiile pe care cele două tipuri de cercetare le pot îndeplini, deosebit de utile în considerarea rolului surselor de dovezi ale validității:

1. *Cercetarea reprezentării constructului se ocupă de identificarea diferențelor dintre sarcinile cerute de test, iar cercetarea domeniilor nomotetice se ocupă de diferențele dintre persoanele testate.* În cercetarea reprezentării constructului, un proces, o strategie sau o arie de cunoștințe identificate prin descompunerea sarcinii (codarea fonetică, raționamentul secvențial sau capacitatea de a înțelege texte de nivel elementar) pot fi considerate esențiale pentru performanța la o sarcină a testului, fără să ducă la diferențe sistematice în populația de subiecți alcătuită din cititori. Pe de altă parte, pentru a investiga domeniul nomotetic al scorurilor obținute la test (i.e. rețeaua de relații dintre ele și alte instrumente) este necesar să dispunem de date ale diferențelor individuale și ale variabilității lor la persoanele testate. Acest fapt întărește importanța crucială pe care o are variabilitatea scorurilor pentru a obține informații ce pot fi utilizate pentru a face afirmații despre o persoană sau a lua decizii cu privire la oameni, discutate anterior în Capitolele 2 și 3. Dacă scorurile unui grup de persoane la un test destinat să evalueze aptitudinea de a înțelege texte de nivel elementar, de exemplu, urmează să fie corelate cu altceva — sau utilizate pentru a determina altceva decât faptul că acești oameni, ca grup, posedă acea aptitudine — trebuie să existe o anumită variabilitate a scorurilor.
2. *Validarea aspectelor sarcinilor cerute de test din reprezentarea constructului este independentă de dovezile favorabile care pot fi acumulate în termenii domeniului nomotetic al scorurilor testului și viceversa.* Cu alte cuvinte, deși am putea cunoaște precis procesele implicate în performanța la sarcinile testului, în absența corelațiilor semnificative cu comportamente sau cu alte instrumente semnificative, scorurile obținute la test pot avea o utilitate limitată. Din același motiv, este posibil să obținem o rețea puternică de relații între scorurile la test și alte instrumente fără a avea o idee clară a constructului pe care-l reprezintă aceste scoruri. De exemplu, Embretson (1983) amintește cazul scorurilor la testele de inteligență care au un domeniu nomotetic puternic (corelează mai mult sau mai puțin puternic cu diferite alte instrumente), dar au baze teoretice relativ neclare.

Schema conceptuală trasată de Embretson (1983) menține noțiunea de validare a constructului drept mod unitar și comprehensiv

Nu uitați!

Oamenii fac tot timpul inferențe pe baza observațiilor sau eșantioanelor de comportament. De exemplu, dacă auzim pe cineva vorbind și făcând greșeli gramaticale, putem deduce că acea persoană este needucată. Dacă o persoană vine la timp la întâlniri invariabil, putem deduce că este punctuală. Unele dintre inferențele noastre sunt corecte, altele nu. Unele contează, altele nu. Dacă inferențele pe care le facem contează suficient de mult pentru noi pentru a dori să le stabilim corectitudinea, trebuie

1. să ne definim termenii neechivoc (Ce înțelegem prin „needucat"? „A veni la timp invariabil" surprinde întregul concept de punctualitate?)
2. să investigăm fidelitatea observațiilor noastre (Acea persoană face întotdeauna greșeli gramaticale sau doar în anumite circumstanțe? Prietenul nostru vine la timp la toate întâlnirile sau doar la cele pe care am avut ocazia să le observăm?) și
3. să decidem dacă există dovezi suficiente pentru a justifica inferențele pe care dorim să le facem, pe baza definițiilor noastre și a datelor disponibile (A fi la timp la toate întâlnirile *este* suficient pentru a judeca punctualitatea unei persoane) sau dacă trebuie să coroborăm inferențele cu date suplimentare (persoana prezintă și alți indicatori pentru ceea ce înțelegem prin „needucat"?).

Testele psihologice sunt instrumente destinate să ajute la rafinarea și cuantificarea observațiilor comportamentale în scopul de a face inferențe despre indivizi, grupuri sau constructe psihologice. În mod fundamental, scorurile obținute la un test psihologic sunt valide în măsura în care ne pot ajuta să facem inferențe exacte.

de exprimare a abordării științifice a integrării oricărei dovezi inportante pentru semnificația sau interpretarea scorurilor unui test. În același timp, oferă o bază pentru a face distincția între (a) sursele de dovezi pentru validitatea scorurilor testului care sunt

relevante pentru a cunoaște ce măsurăm (i.e. reprezentarea constructului) și (b) cele care au de-a face în principal cu inferențele pe care le putem face pe baza a ceea ce măsurăm (i.e. domeniul nomotetic). Trebuie observat că aceste surse de dovezi pot fi, și sunt deseori interrelaționate și ambele presupun elemente observabile și teoretice, precum și modele sau postulate care privesc interrelațiile dintre elemente.

Surse de dovezi ale validității

În general, esența judecăților cu privire la validitatea scorurilor obținute la test se focalizează pe relația dintre ceea ce reprezintă scorurile și întrebările la care utilizatorii testului trebuie să răspundă prin folosirea testelor. Întrebările pe care le punem determină tipul de dovezi de care avem nevoie, precum și relațiile logice — inducție și deducție — necesare pentru a ne adresa următoarelor aspecte (a) ce măsurăm prin intermediul testelor și (b) ce inferențe se pot face ponind de la scorurile testului. De asemenea, trebuie să fie clar în acest punct că, pe măsură ce semnificația sau impactul potențial al răspunsurilor este mai mare, dovezile trebuie să fie mai convingătoare. În continuare ne vom ocupa de tipurile de dovezi necesare pentru a valida inferențele făcute pe baza scorurilor unui test, înțelegând faptul că interpretarea propusă pentru scorurile obținute la un test determină cadrul conceptual pentru validarea sa. Tabelul 5.1 prezintă o listă a categoriilor majore în care pot fi clasificate aspectele validității alături de sursele principale de dovezi pentru fiecare dintre ele, discutate în restul acestui capitol. Este important să recunoaștem de la început faptul că nici aspectele validității, nici sursele sau tipurile de dovezi asociate cu acestea nu se exclud reciproc. De

fapt, strategiile de validare ar trebui să fie încorporate ca surse numeroase de dovezi la fel de utile sau la fel de adecvate pentru scopurile unui test.

Validitatea bazată pe dovezi cu privire la conținutul testului și procesul de formulare a răspunsului la test

Unele teste psihologice sunt proiectate pentru a eșantiona comportamente care pot fi legate mai mult sau mai puțin direct de inferențele pe care dorim să le facem pe baza scorurilor lor. Pe larg, aceste instrumente intră în categoria testelor cu raportare la criteriu sau la conținut, discutate deja pe larg în Capitolul 3. Așa cum notam acolo, cea mai mare parte a acestor teste este utilizată în cadrul educațional și ocupațional, deși ele pot fi aplicate și în domenii (evaluarea neuropsihologică) în care este necesar să se stabilească dacă o persoană este sau nu capabilă să îndeplinească o anumită sarcină ce poate avea semnificație diagnostică. Aceste teste sunt alcătuite din itemi care eșantionează cunoștințele dintr-un domeniu de conținut bine determinat sau cer subiecților testați să demonstreze că ei posedă o aptitudine dată sau o competență într-o anumită abilitate. Procedurile de validare pentru testele de acest tip sunt din multe puncte de vedere cele mai simple și cele mai acceptate aspecte ale construirii testelor, datorită faptului că dovezile pe baza cărora se fac astfel de inferențe pot fi susținute logic și prin relații demonstrabile între conținutul testului și constructul pe care testul este destinat să-l reprezinte.

Dovezile validității scorurilor testului care derivă din conținutul testului pot alcătui un instrument nou încă de la început, prin alegerea itemilor sau sarcinilor care sunt incluse în test.

Tabelul 5.1 Aspecte ale validității de construct și sursele de dovezi legate de aceasta

Aspecte ale validității de construct	Surse de dovezi[a]
În raport cu conținutul	Relevanță și reprezentativitate a conținutului testului și a proceselor implicate în răspunsul la sarcină
	Validitate de aspect (i.e. aparență superficială)
Tipare de convergență și divergență	Consistență internă a rezultatelor testelor și alte măsuri ale fidelității
	Corelații între teste și subteste
	Matrice multi-trăsătură–multi-metodă
	Diferențierea scorurilor consonantă cu diferențele așteptate pe baza vârstei și a altor variabile de statut
	Rezultate experimentale (i.e. corespondență între scorurile testului și efectele prezise ale intervențiilor experimentale sau ipotezele bazate pe teorie)
	Analiză factorială exploratorie
	Echivalarea structurală a tehnicilor de modelare
În raport cu criteriul	Precizia deciziilor bazate pe validarea concurentă (i.e. corelații între scorurile testelor și criteriile existente)
	Precizia deciziilor sau predicțiilor bazate pe validarea predictivă (i.e. corelații dintre scorurile testului și predicția făcută pentru criteriu)

[a] Vezi textul din Capitolul 5 pentru explicația termenilor

Cerința primară pentru a construi teste de acest tip este specificarea foarte atentă a domeniului de conținut, a proceselor cognitive, a aptitudinilor sau tipurilor de performanță ce urmează a fi eșantionate de test și a importanței sau ponderii lor relative.

- În context educațional, exemple de astfel de specificări pot fi găsite în curriculele școlare, în planul de curs, în manuale și orice alte materiale care subliniază, definesc sau stabilesc priorități ale obiectivelor experienței educaționale sau de formare atât în termenii cunoașterii conținutului, cât și în cei ai capacității de performanță. Procesul de delimitare a domeniilor de cunoștințe și de determinare a rezultatelor dorite ale instruirii ține de profesori, formatori și alți experți care determină curricula sau scriu manuale pentru diferite discipline.
- În cadrul ocupațional, specificarea domeniilor de abilități sau cunoștințe ce urmează a fi eșantionate de test se bazează pe analiza muncii. *Analiza muncii* se referă la oricare dintre diferitele metode ce urmăresc descoperirea naturii unei munci date prin descrierea elementelor, activităților, sarcinilor și îndatoririlor legate de ea (vezi Brannick & Levine, 2002). Metodologia de analiză a muncii are diferite aplicații în managementul resurselor umane, incluzând evaluarea performanțelor și, printre altele, determinarea nevoilor de formare. În contextul selecției și clasificării ocupaționale, analiza muncii — bazată pe date obținute de la angajatori, supervizori și/sau angajați curenți — este utilizată pentru a indica precis aptitudinile, abilitățile și bagajul de cunoștințe necesare pentru performanța în muncă.

- În evaluarea neuropsihologică, specificarea proceselor și capacităților cognitive ce urmează a fi evaluate derivă din cunoștințele empirice și teoretice despre legăturile dintre sistemul nervos central și funcțiile comportamentale. Natura conținutului evaluării neuropsihologice se bazează pe dovezi științifice și clinice acumulate despre relațiile creier–comportament.

Info 3.7 din capitolul 3 enumeră câteva exemple simple de obiective și itemi tipici ai testelor cu raportare la domeniu. Exemple mai extinse, inclusiv tabele de specificații pentru testele cu raportare la conținut și ghiduri ale modului în care sunt realizate pot fi găsite în Gronlund (2003) și Linn și Gronlund (1995, pp. 119–125). Indiferent de cadrul în care este aplicat un test cu raportare la conținut, din momentul în care au fost stabilite specificațiile pentru cunoștințe, abilități sau procese ce urmează a fi evaluate prin test, procedurile de validare a conținutului presupun revizuirea critică și examinarea conținutului testului din două perspective. Prima este *relevanța* conținutului eșantionat de test pentru domeniul specificat; a doua este *reprezentativitatea* conținutului eșantionat de test în raport cu specificațiile domeniului pe care este destinat să-l acopere. Deși aspectul relevanței trebuie să se bazeze pe un consens al experților, el poate fi susținut și de rezultate empirice, cum ar fi diferențele dintre scorurile obținute de elevii din clase succesive sau de indivizi aflați în diferite stadii ale procesului de formare. Autorii testelor trebuie să susțină declarațiile privind validitatea scorurilor testelor raportate la conținut în manuale, cărți tehnice și alte astfel de surse de documentație pentru teste. Atunci când baza primară a dovezilor de validare a unui instrument se centrează pe conținutul, abilitățile sau

procesele cognitive specifice pe care le evaluează, este necesară o descriere a procedurilor sistematice utilizate pentru a asigura relevanța și reprezentativitatea conținutului testului pentru domeniile specificate. Info 4.3 prezintă un exemplu al rolului pe care reprezentativitatea neadecvată a acoperirii conținutului îl poate juca în subminarea fidelității și validității scorurilor testului cu raportare la conținut.

Testarea educațională

Scorurile a căror validitate derivă dintr-o legătură directă și demonstrabilă între conținutul testului și specificațiile utilizate în construirea testului sunt numeroase la toate nivelurile de educație și formare în care rezultatele instruirii pot fi definite fără ambiguitate. Aproape toate testele construite de profesori pentru a fi utilizate la clasă intră în această categorie, ca și multe dintre testele standardizate publicate de ETS și ACT și alte organizații similare. Scopul primar al acestor instrumente este de a determina achizițiile educaționale — adică ce au învățat elevii în perioada de școală. Scorurile obținute la aceste teste pot răspunde celor mai directe întrebări, cum ar fi „În ce măsură elevul stăpânește domeniul specificat?" sau „În ce măsură subiectul testat a atins un anumit grad de măiestrie sau de competență în abilitatea aflată în discuție?" Instrumentele cu raportare la domeniu sau la conținut pot fi aplicate în scopul luării unor numeroase decizii, printre care acordarea notelor, oferirea calificativelor în urma unei examinări, conferirea unui grad sau a unei diplome după urmarea unui program de studii, certificarea sau atestarea unor indivizi pentru a practica într-un domeniu dat sau chiar și pentru a stabili dacă o persoană este pregătită să treacă la un nivel mai avansat al formării profesionale. În mod tipic, deciziile bazate pe aceste

teste pivotează în jurul nivelelor de stăpânire a domeniului sau a conținutului prezentate de subiecții testați. Aceste nivele, la rândul lor, pot fi măsurate în termeni de note procentuale, ranguri percentilice în comparație cu grupuri normative adecvate sau determinări simple de tipul admis/respins bazate pe criterii prestabilite, așa cum s-a discutat în Capitolul 3. Info 5.3 enumeră anumite exemple tipice de teste educaționale standardizate, alături de scopurile pentru care au fost construite (i.e. cunoștințele și abilitățile pe care sunt destinate să le evalueze) și aplicațiile lor primare. Exemple de întrebări și descrieri mai elaborate ale acelor teste sunt disponibile pe site-uri de pe internet enumerate în Info 5.3.

Testarea ocupațională

Multe instrumente utilizate pentru a selecta sau repartiza candidații la un loc de muncă constau din eșantioane sau simulări ale activității care cer, de fapt, să se realizeze sarcinile din care este alcătuită activitatea pe care urmează să o desfășoare la locul de muncă (teste de dactilografiere) sau eșantioane de comportament care pot fi legate direct de performanța în muncă prin intermediul analizei muncii. Unele dintre aceste teste sunt construite în interior, de către angajatori și utilizează norme sau criterii de performanță locale. Altele sunt instrumente standardizate care oferă scoruri normative pentru indivizii din diferite ocupații și măsoară constructe cu lărgimi diferite. Numeroase exemple de astfel de teste pot fi găsite în secțiunea Vocations din Classified Subject Index în *Tests in Print VI* (Murphy et al., 2002). Două dintre acestea pot fi utilizate pentru a ilustra diversitatea testelor de acest tip și sunt descrise în Info 5.4.

Info 5.3

Exemple de teste educaționale standardizate care utilizează ca sursă principală de validare dovezile bazate pe conținut

Titlul testului	Scopul principal al testului	Aplicație primară	Adrese pe internet ale descrierilor testului și exemplelor din test
Test of English as a Foreign Language (TOEFL)	Evaluează cunoștințele de limbă engleză ale persoanelor a căror limbă maternă nu este limba engleză	Determină dacă studenții străini posedă cunoștințe de engleză suficiente pentru a fi admiși la colegiu	http://www.toefl.org
College-Level Examination Program (CLEP) Introductory Psychology Test	Măsoară cunoașterea materialului predat în mod obișnuit într-un curs de introducere în psihologie cu durata de un semestru	Determină dacă studenții stăpânesc suficient de bine psihologia introductivă pentru a li se acorda prin examinare un credit.	http://www.collegeboard.com/clep

În parte din cauza costurilor și dificultăților pe care le presupune dezvoltarea și validarea instrumentelor de evaluare a abilităților la nivel local, organizația ACT (anterior American College Testing Program) (www.act.org) a inițiat un program numit sistemul WorkKeys, care combină un număr de componente ce urmăresc să ajute întreprinderile să recruteze, să selecteze, să angajeze

Titlul testului	Scopul principal al testului	Aplicație primară	Adrese pe Internet ale descrierilor testului și exemplelor din test
ACT Assessment Science Reasoning Test	Măsoară interpretarea, analiza, evaluarea, raționamentul și abilitățile de rezolvare de probleme cerute în științele naturale, incluzând biologie, chimie, fizică și științe ale pământului și spațiale.	Evaluarea cunoștințelor și abilităților pe care un student le-a achiziționat pentru a determina dacă este gata să desfășoare o activitate de nivelul colegiului.	http://www.act.org
National Assessment of Educational Progress (NAEP)	Măsoară cunoașterea materiei și abilitățile în citire, matematică, știință, scriere, istoria SUA, cultură civică, geografie și arte	Oferă informații despre performanța populației de studenți și a subgrupurilor din Statele Unite și din statele participante	http://nces.ed.gov

și să își instruiască angajații. Componenta „profilul postului" permite deținătorilor postului sau supervizorilor lor, în colaborare cu experții ACT să selecteze cele mai importante sarcini pentru un post dat și să identifice abilitățile și nivelul acestora necesare pentru o performanță ridicată în acea muncă. Aspectul de evaluare al WorkKeys oferă instrumente standardizate pentru a evalua

nivelele de abilitate ale candidaților sau angajaților în câteva arii importante, cum ar fi, printre altele, Tehnologie aplicată, Scrisori de afaceri, Localizarea informației, Ascultare și Munca în echipă. De exemplu, evaluarea abilităților de Localizare a informației oferă întrebări cu patru nivele de complexitate din ce în ce mai ridicată și este proiectată pentru a măsura abilități care pornesc de la găsirea informației incluse în grafice elementare de la locul de muncă — cum ar fi formulare simple, grafice cu coloane și planul etajului — și ajung până la concluzii trase pornind de la informații prezentate în diagrame, tabele, modele foarte detaliate. Pe baza comparării informației oferite de aceste instrumente de evaluare a abilităților și a nivelelor minime ale unei abilități cerute de profilul postului, angajatorii pot evalua calificarea candidaților sau nevoile de instruire ale angajaților actuali.

Dovezi ale validității de conținut în alte contexte de evaluare

Măsura în care itemii testului sunt relevanți și reprezentativi pentru un construct poate fi o sursă suplimentară de date ale validității pentru instrumente din aproape orice domeniu. De exemplu:

- *În evaluarea neuropsihologică*, așa cum am menționat anterior, specificațiile cu privire la procesele cognitive și capacitățile comportamentale ce urmează a fi evaluate sunt derivate din cunoștințe teoretice și empirice bine stabilite despre relațiile dintre funcțiile cognitive și comportamentale și fundamentul neurologic ce se presupune a se afla la baza acestora. Astfel, într-o mare măsură, procesul de evaluare neuropsihologică se bazează pe cunoașterea de către experți a dovezilor acumulate științific cu privire la relațiile creier-comportament. O baterie neuropsihologică adecvată trebuie să includă itemi care investighează o arie de comportamente suficient de largă și reprezentativă încât să aducă dovezi ale capacității funcționale sau deteriorării diferitelor sisteme pe care este destinată să le măsoare (vezi Franzen, 2000; Lezak, 1995). De exemplu, Boston Diagnostic Aphasia Examination (Goodglass, Kaplan, & Barresi, 2001) oferă o eșantionare sistematică a câtorva funcții de comunicare, cum ar fi înțelegerea auditivă și exprimarea orală, pentru a ajuta în diagnosticarea sindroamelor afazice și a tulburărilor de limbaj.
- *În evaluarea personalității* multe instrumente cu autoraportare — cum ar fi listele, inventarele și chestionarele de atitudini sau de opinii — se bazează într-o mare măsură pe conținutul itemilor lor pentru susținere în generarea ipotezelor sau construirea inferențelor despre constructele particulare pe care urmăresc să le evalueze. Procedurile structurate de observare și diferite inventare utilizate pentru a colecta date bazate pe rapoarte ale colegilor, părinților, partenerilor, profesorilor sau ale altor observatori utilizează, de asemenea, conținutul itemilor ca sursă primară de dovezi despre validitate. În mod similar, testele psihologice destinate să ajute în diagnosticarea tulburărilor psihice includ deseori itemi care reflectă aspecte simptomatice critice ale sindroamelor pe care au fost proiectate să le diagnosticheze sau chiar pot fi alcătuite în întregime din astfel de itemi. Exemplele de teste de acest tip includ Beck Depression Inventory (BDI), State-Trait Anxiety Inventory (STAI), Symptom Checklist-90-Revised (SCL-90-R) și Attitudes Toward Women Scale (AWS; Spence & Helmreich, 1972).

Info 5.4

Exemple de teste ocupaționale standardizate care utilizează dovezile bazate pe conținut ca sursă de validare

Titlul testului	Constructul evaluat	Descriere	Aplicație principală
Craeford Small Parts Dexterity test (CSPDT)[a]	Coordonare ochi-mână și dexteritate motorie fină	Constă din două sarcini: (a) utilizarea unor pensete pentru a introduce ace mici în găurile unei plăci și apoi pentru a plasa mici inele pe ace; (b) fixarea unor șuruburi pe o placă și strângerea lor cu o surubelniță. Factorul major în scorarea acestui test este viteza de îndeplinire a sarcinii.	Este utilizat pentru a determina dacă un individ are dexteritatea manuală necesară pentru o poziție care presupune o activitate precisă cu mâinile, cum ar fi gravura sau repararea ceasurilor.

Dovezi ale validității din perspectiva subiecților testați

Relevanța și reprezentativitatea conținutului testului sunt de asemenea legate de o problemă care este mai puțin concretă decât validitatea scorurilor, dar, cu toate acestea, destul de importantă. *Validitatea de aspect (de fațadă)* se referă la ceea ce, în aparență, măsoară testul, din perspectiva subiectului testat sau a altui observator naiv. Toate instrumentele discutate până în acest punct aveau o anumită validitate de aspect atunci când erau utilizate

Titlul testului	Constructul evaluat	Descriere	Aplicație principală
Clerical Abilities Battery (CAB)[a]	Câteva componente ale unei largi game de activități identificate prin analiza comportamentelor generale ale funcționarilor.	CAB cuprinde câteva subteste: Completare, Comparare, Copierea informației, Utilizarea tabelelor, Corectură, Abilități matematice primare și Raționament aritmetic.	Utilizat pentru rectutarea și evaluarea funcționarilor. Într-o trecere în revistă a Mental Measurement Yearbook, Randhawa (1992) afirmă că eșantionarea și gama de sarcini din subtestele CAB nu este suficient de reprezentativă și sugerează că este nevoie de date suplimentare pentru standardizare, fidelitate și validitate predictivă. Cu toate acestea, admite faptul că procesul de construire și formatul bateriei sunt impresionante și oferă bazele unui instrument cu mare potențial.

[a] Publicat de Psychological Corporation (http://www.PsychCorp.com)

în contextele pe care le-am discutat. Ele le păreau subiecților a fi consonante cu scopurile educaționale, ocupaționale, clinice sau de investigare declarate din situațiile de testare în care sunt aplicate de obicei. Chiar dacă validitatea de aspect nu este în mod

necesar un indicator al validității din perspectivă psihometrică, fără îndoială este o trăsătură dezirabilă, deoarece promovează o anumită raportare a subiectului la testare, acceptarea testării și a rezultatelor testului. În condițiile în care subiecților testați li se pare că un test are un conținut neadecvat sau irelevant, este probabil ca disponibilitatea lor de a coopera la procesul testării să fie subminată. Astfel, autorii testelor trebuie să ia în considerare aparența de validitate din perspectiva tuturor părților — inclusiv cea a subiecților sau a altor neprofesioniști — și, ori de câte ori este posibil, să încorporeze în test un conținut care pare relevant și adecvat pentru situațiile în care testul va fi utilizat.

Dovezi ale validității bazate pe explorarea tiparelor de convergență și divergență

Pe măsură ce interpretarea scorurilor trece de relațiile directe și lipsite de echivoc dintre conținutul testului și volumul de cunoștințe, abilitățile și procesele funcționale pe care testele sunt destinate să le evalueze, ea începe să se bazeze din ce în ce mai mult pe surse indirecte de dovezi ale validității. Această afirmație este adevărată mai ales pentru testele din aria personalității, nu numai deoarece constructele pe care le evaluează sunt de obicei mai teoretice și mai abstracte decât cele evaluate de testele cognitive, ci și pentru că răspunsurile subiecților testați la instrumentele de evaluare a personalității sunt influențate de mult mai mulți determinanți stilistici și situaționali decât răspunsurile la un test cognitiv.

Există un număr mare și în continuă creștere de metode care pot fi utilizate pentru a amplifica semnificația scorurilor unui test dincolo de relevanța și reprezentativitatea conținutului testului. Numitorul comun al tuturor acestor procedee este faptul că

produc dovezi sub forma tiparelor de convergență și divergență între scorurile testului și alte variabile (vezi Tabelul 5.1). Deși o explicație detaliată a acestor metode depășește obiectivele acestei cărți, putem face o descriere elementară a celor mai întâlnite procedee.

Fidelitatea scorurilor ca sursă de dovezi pentru validitate

Investigațiile asupra fidelității scorurilor unui test din punctul de vedere al stabilității, al diferențelor inter-evaluatori, al erorii de eșantionare a conținutului și al eterogenității conținutului oferă dovezi cu privire la coeziunea sau specificitatea conținutului testului. Așa cum am discutat în Capitolul 4, fidelitatea scorurilor poate fi văzută ca dovadă preliminară a obținerii unui instrument de măsurare a unui eșantion de comportament în care se poate avea încredere și astfel poate oferi dovezi indirecte pentru validitatea scorurilor testului. Dacă, de exemplu, un test este construit pentru a evalua un construct unidimensional cum ar fi abilitatea de a scrie corect, coeficienți mari de consistență internă ar putea susține afirmația de unidimensionalitate. În mod similar, dacă se poate obține o consistență a scorurilor pentru diferiți evaluatori, se poate presupune că toți au utilizat același criteriu și, astfel, au evaluat probabil aceleași caracteristici. Presupunând stabilitatea constructelor evaluate — de exemplu, o trăsătură sau un tip de personalitate — o condiție preliminară esențială a dovezilor validității vor fi scoruri înalte ale fidelității test-retest.

Corelații între teste și subteste

O modalitate simplă și frecvent utilizată de a acumula dovezi că un anumit test măsoară constructul pe care își propune să-l măsoare este stabilirea unor corelații înalte între scorurile sale și cele ale altor instrumente destinate să evalueze același construct.

Unul dintre cele mai simple exemple ale acestui tip de procedură apare atunci când testele sunt revizuite și renormate. În astfel de cazuri, manualele testului prezintă aproape invariabil corelațiile înalte dintre edițiile noi și cele vechi drept dovezi ale faptului că ambele variante măsoară aceleași constructe. Este oarecum similar cu calcularea corelațiilor dintre formele alternative ale unui test pentru a stabili fidelitatea sau consistența internă a scorurilor diferitelor forme ale testului. Cu toate acestea, ne putem aminti din Capitolul 3 că, deși corelațiile dintre versiunile vechi și cele revizuite ale testelor sunt foarte înalte, scorurile normative pentru versiunile restandardizate tind să alunece într-o direcție sau alta datorită schimbărilor din populație după diferite perioade de timp.

Similar, autorii testelelor prezintă în mod tipic corelații între scorurile testelor lor și cele ale instrumentelor comparabile drept dovadă a validității scorurilor. De exemplu, toate manualele scalelor importante de inteligență prezintă corelații între scorurile acestor scale și cele ale altor instrumente bine consolidate de același tip. Prin examinarea acestor date se poate afla — de exemplu — că între WAIS-III Full Scale IQ și scorul compus global la Stanford-Binet-IV (SB-IV) calculat pentru un eșantion de 26 de indivizi care au răspuns la ambele teste s-a obținut o corelație de 0,88 (Psychological Corporation, 1997, p. 84) sau că a existat o corelație de 0,87 între scorurile compuse la SB-IV și la Kaufman Adolescent & Adult Intelligence Test (KAIT), pentru un eșantion alcătuit din 72 de indivizi testați cu ambele instrumente (Kaufman & Kaufman, 1993, p. 100). Coeficienții de corelație cu astfel de valori sunt tipici pentru scalele de inteligență majore și servesc pentru a corobora faptul că o mare parte din varianța scorurilor obținute la aceste teste este comună.

Pot fi obținuți coeficienți de corelație și între scorurile la subtestele din diferite scale. Exemple tipice ar putea fi corelații între scorurile diferitelor scale de depresie, de exemplu scorurile la scala Depresie din MMPI-2 și cele ale scalei Distimie (r =0,68) din Millon Clinical Multiaxial Inventory-III (MCMI-III), sau scorurile la Beck Depression Inventory și cele la scala Depresie majoră MCMI-III (r = 0,71; Millon, Millon, & Davis, 1994, pp. 126, 129). Așa cum era de așteptat, coeficienții de corelație calculați între diferite tipuri de teste și subteste sunt prezenți din plin în manualele testelor și în literatura de specialitate, deși, de cele mai multe ori, acești indicatori nu sunt foarte convingători și nu aduc multe informații.

Corelațiile dintre teste sunt atât de prezente deoarece datele pentru studiile corelaționale pe eșantioane mici sunt ușor de obținut, mai ales în cazul testelor tip creion-hârtie care pot fi administrate cu ușurință în grup. Corelațiile obținute astfel pot, bineînțeles, să se întindă între zero și ±1,00, în funcție de scorurile respective și de natura eșantioanelor utilizate (cf. Capitolul 2, mai ales secțiunea despre restricțiile de interval și corelație). Deși semnificația oricărui coeficient obținut poate fi interpretată, dacă se acumulează suficiente date care arată corelații consecvent puternice sau consecvent slabe între instrumente, pot fi stabilite anumite tipare de convergență și divergență. Aceste tipare informează utilizatorii testului despre gradele aproximative de varianță comună dintre seturile de scoruri și, în mod indirect, despre semnificația scorurilor. Corelații consecvent înalte între instrumentele construite pentru a evalua un construct dat — cum ar fi corelațiile indicate în paragraful precedent pentru scalele de depresie — pot fi considerate drept dovezi ale *validității convergente*, adică dovezi cu privire la similaritatea sau identitatea constructelor pe care acestea le evaluează. Din aceleași motive, dovezile *validității*

Unul dintre cele mai simple exemple ale acestui tip de procedură apare atunci când testele sunt revizuite și renormate. În astfel de cazuri, manualele testului prezintă aproape invariabil corelațiile înalte dintre edițiile noi și cele vechi drept dovezi ale faptului că ambele variante măsoară aceleași constructe. Este oarecum similar cu calcularea corelațiilor dintre formele alternative ale unui test pentru a stabili fidelitatea sau consistența internă a scorurilor diferitelor forme ale testului. Cu toate acestea, ne putem aminti din Capitolul 3 că, deși corelațiile dintre versiunile vechi și cele revizuite ale testelor sunt foarte înalte, scorurile normative pentru versiunile restandardizate tind să alunece într-o direcție sau alta datorită schimbărilor din populație după diferite perioade de timp.

Similar, autorii testelelor prezintă în mod tipic corelații între scorurile testelor lor și cele ale instrumentelor comparabile drept dovadă a validității scorurilor. De exemplu, toate manualele scalelor importante de inteligență prezintă corelații între scorurile acestor scale și cele ale altor instrumente bine consolidate de același tip. Prin examinarea acestor date se poate afla — de exemplu — că între WAIS-III Full Scale IQ și scorul compus global la Stanford-Binet-IV (SB-IV) calculat pentru un eșantion de 26 de indivizi care au răspuns la ambele teste s-a obținut o corelație de 0,88 (Psychological Corporation, 1997, p. 84) sau că a existat o corelație de 0,87 între scorurile compuse la SB-IV și la Kaufman Adolescent & Adult Intelligence Test (KAIT), pentru un eșantion alcătuit din 72 de indivizi testați cu ambele instrumente (Kaufman & Kaufman, 1993, p. 100). Coeficienții de corelație cu astfel de valori sunt tipici pentru scalele de inteligență majore și servesc pentru a corobora faptul că o mare parte din varianța scorurilor obținute la aceste teste este comună.

Pot fi obținuți coeficienți de corelație și între scorurile la subtestele din diferite scale. Exemple tipice ar putea fi corelații între scorurile diferitelor scale de depresie, de exemplu scorurile la scala Depresie din MMPI-2 și cele ale scalei Distimie (r =0,68) din Millon Clinical Multiaxial Inventory-III (MCMI-III), sau scorurile la Beck Depression Inventory și cele la scala Depresie majoră MCMI-III (r = 0,71; Millon, Millon, & Davis, 1994, pp. 126, 129). Așa cum era de așteptat, coeficienții de corelație calculați între diferite tipuri de teste și subteste sunt prezenți din plin în manualele testelor și în literatura de specialitate, deși, de cele mai multe ori, acești indicatori nu sunt foarte convingători și nu aduc multe informații.

Corelațiile dintre teste sunt atât de prezente deoarece datele pentru studiile corelaționale pe eșantioane mici sunt ușor de obținut, mai ales în cazul testelor tip creion-hârtie care pot fi administrate cu ușurință în grup. Corelațiile obținute astfel pot, bineînțeles, să se întindă între zero și ±1,00, în funcție de scorurile respective și de natura eșantioanelor utilizate (cf. Capitolul 2, mai ales secțiunea despre restricțiile de interval și corelație). Deși semnificația oricărui coeficient obținut poate fi interpretată, dacă se acumulează suficiente date care arată corelații consecvent puternice sau consecvent slabe între instrumente, pot fi stabilite anumite tipare de convergență și divergență. Aceste tipare informează utilizatorii testului despre gradele aproximative de varianță comună dintre seturile de scoruri și, în mod indirect, despre semnificația scorurilor. Corelații consecvent înalte între instrumentele construite pentru a evalua un construct dat — cum ar fi corelațiile indicate în paragraful precedent pentru scalele de depresie — pot fi considerate drept dovezi ale *validității convergente*, adică dovezi cu privire la similaritatea sau identitatea constructelor pe care acestea le evaluează. Din aceleași motive, dovezile *validității*

discriminatorie bazate pe corelații consecvent scăzute între instrumente despre care se presupune că diferă, pot fi utilizate pentru a demostra identitatea constructelor pe care le investighează. Un exemplu de astfel de tip de tipar divergent poate fi găsit în corelațiile dintre scorurile la scala Bipolar, Maniacal a MCMI-III și scala Depresie din MMPI-2 ($r = 0,06$), ca și între scorurile la scala Depresie Majoră a MCMI-III și scala Hipomanie din MMPI-2 ($r = 0,08$), ambele calculate pe un eșantion alcătuit din 132 de indivizi (Millon, Millon, & Davis, 1994, pp. 129–130).

Matricea multitrăsătură-multimetodă

În efortul de a organiza colectarea și prezentarea datelor de validare convergentă și discriminatorie, D. T. Campbell și Fiske (1959) au propus un model pe care l-au numit *matricea multitrăsătură-multimetodă (MMTMM)*. Această abordare se referă la o strategie de validare care necesită colectarea datelor pentru două sau mai multe trăsături distincte (anxietate, afiliere și dominanță) prin două sau mai multe metode diferite (chestionare cu autoadministrare, observații ale comportamentului și tehnici proiective). Odată ce aceste date sunt colectate și sunt calculate toate inter-corelațiile dintre ele, pot fi prezentate sub forma unei matrici cum este cea din Tabelul 5.2. Matricile multitrăsătură-multimetodă prezintă (a) coeficienții de fidelitate pentru fiecare instrument, (b) corelațiile dintre scorurile pentru aceeași trăsătură evaluată prin metode diferite (i.e. date ale validității convergente) și (c) corelații între scorurile la diferite trăsături măsurate prin aceleași metode, și (d) între scorurile obținute pentru diferite trăsături evaluate prin metode diferite (ambele constituind date ale validității discriminatorii). Tabelul 5.2 este un MMTMM cu

Tabelul 5.2 O matrice multitrăsătură-multimetodă ipotetică (MMTMM)

Metoda	Trăsătura	Autoadministrare			Observație			Proiectivă		
		Anx	Af	Dom	Anx	Af	Dom	Anx	Af	Dom
Autoadministrare	Anx	((0,90)								
	Af	*0,45*	(0,88)							
	Dom	*0,35*	*0,38*	(0,80)						
Observație	Anx	**0,60**	0,23	0,10	(0,92)					
	Af	0,25	**0,58**	-0,08	*0,47*	(0,93)				
	Dom	0,12	-0,12	**0,55**	*0,30*	*0,32*	(0,86)			
Proiectivă	Anx	**0,56**	0,22	0,11	**0,65**	0,40	0,31	(0,94)		
	Af	0,23	**0,57**	0,05	0,38	**0,70**	0,29	*0,44*	(0,89)	
	Dom	0,13	-0,10	**0,53**	0,19	0,26	**0,68**	*0,40*	*0,44*	(0,86)

Notă: Anx = Anxietate; Af = Afiliere; Dom = Dominanță. Coeficienții de fidelitate sunt scriși în paranteză, pe diagonala principală. Coeficienții de validitate (aceeași trăsătură evaluată prin metode diferite) sunt scriși cu caractere aldine. Toți ceilalți coeficienți sunt indici ai validității discriminatorii a scorurilor obținute pentru diferite trăsături evaluate printr-o singură metodă (reprezentând metoda comună de variantă și fiind trecuți cu caractere italice) și diferite trăsături evaluate prin diferite metode (cu caractere normale).

un tipar al rezultatelor care ar putea fi considerat exemplar pentru acest tip de model de validare. Matricea din acest tabel arată:

- Cei mai înalți coeficienți, care indică un scor al fidelității adecvat (în paranteze) pe diagonala principală;
- Următorii coeficienți ca mărime — cu caractere **aldine** — între măsurări ale **aceleiași** trăsături realizate prin metode **diferite**, indicând convergența dintre scorurile lor;

- Următorii coeficienți ca mărime — cu caractere *italice* — între măsurătorile unor trăsături *diferite* realizate prin *aceeași* metodă, indicând faptul că un mare grad al variației scorurilor se datorează metodelor utilizate
- Cei mai mici coeficienți — cu caractere normale — între măsurători ale trăsăturilor diferite evaluate prin metode diferite indică faptul că instrumentele discriminează destul de bine

Modelul MMTMM este ideal pentru a investiga tiparele de convergență și divergență între scorurile obținute la un test și datele acumulate cu alte tipuri de instrumente de evaluare. Cu toate acestea, reprezintă un standard de validare destul de precis, care este deseori dificil de obținut, mai ales în cazul instrumentelor de evaluare a personalității, ale căror scoruri sunt predispuse să prezinte într-o mare măsură o *variantă de metodă* (i.e. variabilitate care este legată de caracteristicile inerente în metodologiile lor). În plus, modelul MMTMM nu se aplică în forma completă foarte frecvent, deoarece colectarea informațiilor prin metode multiple este destul de laborioasă (vezi Terrill, Friedman, Gottschalk, & Haaga, 2002). Cu toate acestea, variante mai simple ale schemei MMTMM, bazate pe scoruri la testele care măsoară atât constructe similare, cât și constructe nesimilare, deși prin metode similare, sunt implicate din ce în ce mai mult în procesul de validare a unui test. Mai mult, unele instrumente au trăsături care facilitează colectarea datelor din diferite surse ce pot fi utilizate apoi pentru a studia tiparele de convergență și discriminare. De exemplu, varianta revizuită pentru NEO Personality Inventory (NEO PI-R) oferă forme paralele pentru aceleași seturi de itemi — Forma S pentru auto-administrare, și Forma R pentru evaluări făcute de

observatori cum ar fi colegi sau parteneri — care pot fi utilizate pentru a corela și compara scorurile derivate din ambele surse (Costa & McCrae, 1992, pp. 48–50).

Diferențieri în funcție de vârstă

Rezultatele obținute la test care sunt consonante cu tendințe de dezvoltare bine stabilite în grupurile de vârstă sunt deseori privite drept dovezi ale validității scorurilor. De fapt, criteriul diferențierii în funcție de vârstă este una dintre primele surse de dovezi pentru validarea testelor de aptitudine. Ne putem aminti din Capitolul 1 că succesul scalelor inițiale Binet-Simon a fost determinat în principal prin studii care dovedeau că eșantionarea funcțiilor cognitive a produs rezultate care puteau fi utilizate pentru a descrie cantitativ nivelurile de aptitudine a copiilor în termenii nivelelor de vârstă cărora le corespundea performanța la test. În cazul celor mai multe teste de aptitudine, performanța copiilor și a adolescenților din eșantioanele normative prezintă în mod tipic o tendință ascendentă pentru vârstele succesive cronologic. La celălalt capăt al spectrului de vârstă, scăderea performanțelor este observată în eșantioane de adulți vârstnici la instrumente care măsoară aptitudini ce tind să se diminueze cu vârsta, cum ar fi testele de memorie și testele care evaluează rapiditatea performanței. Diferențierea în funcție de vârstă este evidentă și în studiile atent realizate ale tendințelor pe o lungă durată de timp ale performanței indivizilor de diferite vârste la testele de aptitudini mentale, cum ar fi Seattle Longitudinal Study (Schaie, 1994). Creșterile sau descreșterile scorurilor consonante cu așteptările conform vârstei oferă dovezi necesare, dar nu suficiente pentru a arăta că un test măsoară constructul pe care a fost construit să-l măsoare.

Rezultate experimentale

O altă sursă indirectă de dovezi care pot fi utile în validarea scorurilor unui test este oferită de investigații care utilizează scorurile obținute la un test psihologic ca variabilă dependentă pentru a determina efectul intervențiilor experimentale. În aria testării aptitudinilor, aceste dovezi derivă în primul rând din diferențele între scorurile obținute pre- și post- intervenție experimentală, urmărind intervențiile care caută să remedieze deficiențele sau să crească performanța la diferite abilități cognitive și intelectuale. De exemplu, dacă scorurile la un test al dezvoltării conceptuale elementare pentru copii mici (e.g. Bracken Basic Concept Scale-Revised) indică o creștere semnificativă pentru un grup expus la un program de îmbogățire de scurtă durată — schimbarea scorurilor ar putea fi văzută ca dovadă atât a validității lor, cât și a eficacității programului. În mod similar, contrastele pre- și post-intervenție sunt utilizate deseori pentru a documenta validitatea scorurilor obținute la instrumente de evaluare a personalității. Un exemplu de astfel de studiu de validare poate fi găsit în manualul de la Quality of Life Inventory (QOLI; Frisch, 1994, pp. 15–16), care este un instrument pentru evaluarea nivelelor de satisfacție a vieții și a stării de bine subiective ce pot fi utilizate — printre altele — pentru a determina eficiența intervențiilor psihoterapeutice sau de consiliere.

Analiza factorială

O cale de a face față numărului mare de constructe legate de testele existente — și numărului mare de corelații care pot fi obținute din scorurile lor globale, scorurile subtestelor și scorurile itemilor — este dată de o serie de procedee statistice cunoscute colectiv ca *analiză factorială (AF)*. Obiectivul principal al analizelor

factoriale este de a reduce numărul de dimensiuni necesare pentru a descrie datele derivate dintr-un mare număr de măsurători. Este realizată printr-o serie de calcule matematice bazate pe algebră matriceală, destinate să extragă tiparele de inter-corelație dintr-un set de variabile.

Există două moduri fundamentale de a conduce analizele factoriale. Abordarea originală a acestei metode are natură exploratorie și astfel este cunoscută drept *analiză factorială exploratorie* sau AFE; își propune să descopere ce *factori* (i.e. variabile latente sau constructe) sunt subiacenți variabilelor analizate. O a doua abordare, mult mai recentă, este numită *analiză factorială confirmatorie (AFC)* deoarece își propune să testeze ipoteze sau să confirme teorii despre factori a căror existență este presupusă deja. Ambele abordări pot fi utilizate în analizarea datelor unui test psihologic, ca și a numeroase alte tipuri de seturi de date. Analizele confirmatorii sunt mult mai sofisticate din punct de vedere metodologic și vor fi discutate mai târziu în acest capitol ca un subset de tehnici de analiză a structurilor covarianței, cunoscut drept modelul ecuației structurale.

Care sunt pașii analizei factoriale a scorurilor unui test psihologic? Analizele factoriale exploratorii debutează cu o *matrice de corelație*, un tabel care prezintă inter-corelațiile dintre scorurile obținute de un eșantion de indivizi la o gamă largă de teste (sau subteste sau itemi). Faptul că acesta este punctul inițial al AFE este important pentru înțelegerea rezultatelor cercetărilor prin analiză factorială deoarece subliniază două trăsături deosebit de importante ale AF deseori uitate. Ambele trăsături se referă la limitele aplicabilității rezultatelor care apar din AF și anume că rezultatele depind în mare măsură de (a) alegerea instrumentelor incluse în

analiză și (b) componența specifică a eșantionului ale cărui scoruri oferă date pentru analiză.

Tabelul 5.3, Partea A, prezintă un exemplu al unei matrice simple de corelație. Această matrice a fost realizată ponind de la scorurile obținute de 95 de studenți la cinci subteste din Beta III, un test nonverbal de aptitudini intelectuale, apărut din Army Beta (vezi Capitolul 1). Datele reprezintă o parte a unui studiu asupra diferențelor de aptitudine cognitivă determinate de sex (Urbina & Ringby, 2001).

Următorii pași ai analizei factoriale depind de alegerea specifică a tehnicilor utilizate de cercetător. Pot fi utilizate câteva proceduri diferite pentru a conduce aceste analize și pentru extragerea factorilor (vezi Bryant & Yarnold, 1995; Comrey & Lee, 1992). O prezentare a procedeelor analizei factoriale depășește obiectivele pe care ni le-am propus, pentru că ar presupune să intrăm în detalii tehnice, cum ar fi metodele de extragere și rotare a factorilor, care sunt destul de complexe. Cu toate acestea, faptul că există diferite abordări ale AF trebuie menționat și reținut, deoarece se poate ajunge la diferite soluții în funcție de ipotezele și metodele utilizate. Diferențele soluțiilor țin în general de numărul de factori extrași și de independența lor relativă unul față de altul. Pentru înțelegerea acestor aspecte și a altora legate de metodologia AF, vezi Russell (2002). Produsul final al analizelor factoriale este o *matrice factorială*, care este un tabel care prezintă saturațiile fiecăreia dintre variabilele inițiale în factorii extrași prin analize. *Saturațiile factoriale* sunt corelații între instrumentele inițiale din matricea de corelație și factorii care au fost extrași. Partea B a Tabelului 5.3 prezintă matricea factorială pentru cei doi factori extrași prin analiza factorială a datelor din matricea

Tabelul 5.3

A. Matricea de corelație: inter-corelații ale scorurilor obținute la cinci subteste din Beta — III de 95 de studenți

Subtest	Codare	Completare de imagini	Verificare	Imagini absurde	Raționament matriceal
Codare	1,00	0,13	0,62**	0,20	0,05
Completare de imagini		1,00	0,09	0,21*	0,11
Verificare			1,00	0,18	0,20
Imagini absurde				1,00	0,31**
Raționament matriceal					1,00

Notă: Datele au fost preluate din Urbina și Ringby (2001)
*$p \leq 0,05$; **$p \leq 0,01$

B. Matrice factorială pentru cei doi factori extrași din analiza factorială exploratorie (AFE) pentru cinci subteste din Beta-III[a]

Subtest	Saturația în Factorul 1	Saturația în Factorul 2
Codare	**0,90**	0,06
Completare de imagini	0,07	**0,54**
Verificare	**0,88**	0,14
Imagini absurde	0,15	**0,75**
Raționament matriceal	0,02	**0,73**

Notă: Caracterele aldine indică cele mai mari saturații în cei doi factori rotați Varimax.

[a] Factorii 1 și 2 explică 61% din varianța scorurilor subtestelor; 39% din varianță este explicat prin factori specifici fiecărui subtest și prin varianța erorii.

de corelație din Partea A a aceluiași tabel. Această matrice factorială indică faptul că cele două subteste (Codare și Verificare) au saturații foarte mari în Factorul 1 și saturații neglijabile în Factorul 2, în timp ce celelalte trei subteste (Completarea de imagini, Imagini absurde și Raționament matriceal) prezintă un tipar invers al saturațiilor factoriale.

Interpretarea rezultatelor analizelor factoriale. Odată obținute, matricile factoriale pot fi examinate pentru a încerca să se determine natura factorilor care explică cea mai mare parte a varianței setului inițial de date. Factorii sunt apoi etichetați pe baza unei logici inductive. Pentru a identifica și eticheta factorii, se examinează trăsăturile distinctive ale instrumentelor care au cea mai mare și cea mai mică saturație în fiecare factor din matrice. În exemplul nostru, matricea factorială din Tabelul 5.3 sugerează faptul că primul factor presupune *rapiditatea execuției*, deoarece cele două subteste cu cea mai mare saturație în acel factor presupun sarcini extrem de simple și au limite de timp foarte reduse. Al doilea factor are saturații mari în cele trei subteste rămase, toate presupunând probleme care necesită *raționament* pe baza unor stimuli pictoriali sau grafici. Acest tipar al matricei factoriale coincide cu cel derivat din analizele factoriale exploratorii și confirmatorii ale datelor eșantionului de standardizare Beta III, prezentate în manualul Beta III. Factorii 1 și 2 sunt etichetați „Viteza procesării" și „Raționament nonverbal" (Kellogg și Morton, 1999).

Nu uitați!

Factorii nu sunt entități „reale", deși deseori sunt prezentați ca și cum ar fi. Aceștia sunt pur și simplu constructe sau variabile latente care pot fi deduse din tiparele de covarianță dezvăluite prin analize statistice.

Analiza factorială a fost construită de psihologi în încercarea de a investiga bazele subiacente ale interrelației dintre scorurile unui test și în special dintre scorurile diferitelor tipuri de teste de aptitudine. Cu toate acestea, tehnicile AF au fost aplicate neîntârziat datelor testelor de personalitate și descrierilor trăsăturilor de personalitate. Istoria AF, atât în aria aptitudinilor, cât și în cea a personalității, a fost plină de controverse cu privire la adecvarea diferitelor sale metode și la extrapolările care pot fi făcute în mod rezonabil pe baza rezultatelor analizei factoriale (Cowles, 2001, cap. 11).

În domeniul aptitudinilor cognitive, o mare parte a controversei s-a centrat pe factorul general al aptitudinii mentale sau *g* (postulat inițial de Charles Spearman), mai ales prin întrebări legate de semnificația sa și de transmiterea ereditară (Jensen, 1998). Teoretizările și cercetările fundamentale suplimentare din acest domeniu s-au ocupat în principal de aspecte privind natura, numărul și organizarea trăsăturilor intelectuale. O compilație remarcabilă a unei mari părți a literaturii despre analiza factorială a aptitudinilor cognitive, alături de o teorie ierarhică a organizării trăsăturilor cognitive larg acceptată poate fi găsită în cartea scrisă de John Caroll (1993) pe această temă.

În domeniul evaluării personalității, analiza factorială a fost aplicată pentru identificarea și măsurarea dimensiunilor majore necesare pentru o descriere completă a personalității, o temă larg dezbătută. În această arie, au apărut independent două tradiții separate ale cercetărilor prin analiză factorială. Una dintre acestea se centrează încă de la început pe utilizarea datelor din chestionarele de personalitate. Cealaltă — cunoscută drept tradiția *lexicală* — începe reducând numeroasele cuvinte utilizate pentru a descrie atributele personalității la un număr gestionabil de cuvinte prin combinarea sinonimelor. Aceasta a fost urmată de o încercare de a identifica dimensiunile primare ale personalității

prin inter-corelații și analize factoriale ale evaluărilor făcute diferitelor trăsături atribuite unor grupuri eterogene de indivizi de către asociații lor, ca și ale datelor din chestionarele cu autoraportare. Mai recent, cercetările din ambele tradiții s-au unit și au atins un anumit grad de consens. Modelul prevalent se centrează asupra utilizării unui tipar ierarhic de analiză pentru a simplifica colectarea datelor cu diferite grade de generalitate, relevante pentru funcționarea personalității, fiind cunoscut drept *modelul celor cinci factori* (*FFM;* Carroll, 2002; Costa & McCrae, 1992; Digman, 1990; Wiggins & Pincus, 1992).

În ciuda problemelor și limitelor sale, tradiția analizei factoriale a fost extraordinar de fructuoasă pentru testarea psihologică și, mai general, pentru teoretizarea psihologică. Longevitatea acestor metode, ca și rafinarea și extinderea lor continuă, au creat o bogată arhivă de instrumente și date de la care putem să pornim și să continuăm să extindem înțelegerea trăsăturilor și testelor psihologice. Info 5.5 subliniază unele dintre beneficiile AF, cât și limitele sale majore.

Tehnici de modelare a ecuației structurale

Analiza factorială exploratorie este doar unul dintre tipurile de tehnici ale statisticii multivariate care permit cercetătorilor să examineze relațiile dintre mai multe instrumente pentru a încerca să determine constructele subiacente care explică variabilitatea observată. Disponibilitatea din ce în ce mai mare a computerelor puternice și a programelor din ultimele decenii a intensificat ușurința cu care tehnicile de analiză factorială — și alte metode sofisticate de analiză a datelor corelaționale multivariate, cum ar fi analizele multiple ale regresiei — pot fi utilizate pentru a

Info 5.5

Partea I: Beneficii ale analizei factoriale

Validarea constructului: Strângând numeroase instrumente și examinând factorii care par să explice varianța comună a acestor instrumente, putem afla mai multe despre alcătuirea sarcinilor eșantionate de testele psihologice și despre organizarea trăsăturilor în termenii generalității și specificității lor.

Aplicare în practică: atunci când o baterie de teste este alcătuită dintr-un mare număr de teste, rezultatele analizei factoriale oferă un mod de a simplifica interpretarea și de a raporta scorurile subtestelor. Aceasta se realizează prin intermediul scorurilor factorilor, care, esențial, sunt indici care cumulează scorurile subtestelor într-un număr mai redus de categorii coezive de construct derivate din analize factoriale.

Partea a II: Limitele analizei factoriale

Interpretarea rezultatelor oricărui studiu realizat prin analiză factorială nu poate fi extinsă dincolo de datele utilizate în analiză, fie în termenii a ceea ce este măsurat, fie în termenii posibilității generalizării lor la alte populații.

Ce se măsoară? Atât manualul Beta III (un test care își propune să măsoare aptitudinea intelectuală nonverbală) cât și analiza prezentată în Tabelul 5.3 sugerează faptul că cele 5 subteste din Beta III pot fi configurate în doi clusteri care au în comun o mare parte a varianței. Examinarea sarcinilor implicate în cele cinci subteste confirmă faptul că cei doi factori sunt nonverbali. Totuși, aceste date nu pot arăta dacă sau în ce măsură acești factori surprind aspecte esențiale ale aptitudinii intelectuale nonverbale.

Posibilitatea de generalizare a rezultatelor analizei factoriale: datele corelaționale obținute pornind de la 95 de studenți (Urbina & Ringby, 2001) și prezentate în Tabelul 5.3 duc la rezultate similare cu cele obținute în grupul de standardizare Beta III, un eșantion mult mai numeros (N = 1260), și mult mai reprezentativ pentru populația Statelor Unite. Deși această convergență a rezultatelor susține structura factorială obținută în ambele investigații, rămâne deschisă întrebarea dacă această structură se va generaliza la alte populații, cum ar fi oameni din diferite culturi, indivizi cu deficiențe vizuale sau auditive necorectate sau orice alt grup a cărui istorie experiențială diferă semnificativ de cea a eșantioanelor utilizate în cele două analize disponibile.

investiga variabile latente (i.e. constructe) și legăturile cauzale directe și indirecte posibile sau căile de influență dintre acestea.

Un set de procedee care s-a dezvoltat cu rapiditate și care poate fi utilizat pentru a testa plauzibilitatea interrelațiilor presupuse dintre constructe și a relațiilor dintre constructe și instrumentele folosite pentru a le evalua este cunoscut drept *modelarea ecuației structurale (MES)*. Ideea esențială a MES este de a crea unul sau mai multe modele — pe baza teoriilor, rezultatelor anterioare, sau a analizelor exploratorii anterioare — ale relațiilor dintre un set de constructe sau de variabile latente și de a compara structurile de covarianță sau matricile implicate de modele cu matricile de covarianță obținute cu un nou set de date. Cu alte cuvinte, relațiile obținute între date empirice ale unor variabile care măsoară diferite constructe (factori sau variabile latente) sunt comparate cu cele prezise de modele. Corespondența dintre date și modele este evaluată cu ajutorul unor mărimi statistice, numite indici *de potrivire* (goodness of fit) MES oferă câteva avantaje suplimentare față de analizele tradiționale ale regresiei. Avantajele sale derivă în principal din două caracteristici ale acestei metodologii: (a) MES se bazează pe analizele *structurilor de covarianță* (i.e. tipare de covariație între variabile latente sau constructe) ce pot reprezenta influențe directe și indirecte ale variabilelor una asupra celeilalte și (b) MES în mod tipic utilizează indicatori multipli atât pentru variabilele independente, cât și pentru cele dependente din modele și oferă astfel o cale de a explica eroarea de măsură a tuturor variabilelor observate. Cititorii care doresc să urmărească mai detaliat MES și tehnicile legate de aceasta, cum ar fi analiza „path" pot consulta una sau mai multe dintre sursele sugerate în Info 5.6.

În măsura în care este vorba despre validarea unui test psihologic, tehnicile MES sunt utilizate pentru explorarea sistematică a constructelor psihologice și a teoriilor prin cercetări care utilizează testarea psihologică drept o metodă de colectare a datelor pentru unul sau mai mulți indicatori din model. MES poate oferi dovezi care să susțină atât fidelitatea scorurilor unui test, cât și utilitatea acestora ca instrumente pentru unul sau mai multe constructe din model. Cu toate acestea, în prezent, cea mai extinsă aplicare a tehnicilor MES în validarea scorurilor unui test este analiza factorială confirmatorie.

Analizele factoriale confirmatorii (AFC) menționate pe scurt în secțiunea anterioară, presupun specificarea apriori a unuia sau mai multor modele ale relației dintre scorurile testului și factorii sau constructele pe care sunt destinate să le evalueze. În AFC, ca și în

Info 5.6

Surse de informații despre modelarea ecuației structurale (MES)

- Pentru o bună introducere în MES care nu presupune cunoașterea metodelor statistice dincolo de analiza regresiei, vezi:
  Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Erlbaum.
- O prezentare mai avansată a MES care include contribuții ale multor autorități în această metodologie poate fi găsită în:
  Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- Internetul este o sursă excelentă de informații despre diferite subiecte, inclusiv tehnicile MES utilizate un numeroase domenii. Unul dintre cele mai bune puncte de plecare este site-ul creat și menținut de Ed Rigdon, profesor al catedrei de Marketing a Georgia State University. Adresa este: http://www.gsu.edu/~mkteer/sem.html.

toate celelalte tehnici MES, este testată potrivirea dintre direcția și intensitatea interrelațiilor estimate prin diferite modele și rezultatele obținute din datele actuale. Aceste analize au fost facilitate de dezvoltarea programelor pentru computer — cum ar fi LISREL (Jöreskog & Sörbom, 1993) — destinate să genereze valori pentru modelele ipotetice care pot fi testate apoi în raport cu datele actuale.

Exemple de astfel de lucrări sunt din ce în ce mai numeroase atât în literatura de specialitate, cât și în secțiunile despre validitate ale manualelor testelor. Analizele factoriale confirmatorii realizate cu date din eșantionul de standardizare pentru WAIS III (Psychological Corporation, 1997, pp. 106–110) sunt tipice pentru studiile concepute pentru a oferi dovezi ale validității scorurilor testelor psihologice. În aceste analize, au fost evaluate patru modele structurale posibile — un model cu doi factori, cu trei, cu patru și cu cinci factori — și au fost comparate succesiv cu un model general unifactorial pentru a determina care dintre ele oferă cea mai bună potrivire cu datele din eșantionul total și pentru cele mai multe intervale de vârstă din grupului normativ WAIS — III. Rezultatele AFC au indicat că modelul de patru factori oferea cea mai bună soluție globală și confirma tiparele obținute anterior cu AFE a acelorași date. La rândul lor, aceste rezultate au fost folosite ca bază pentru determinarea componenței celor patru indicatori ai scorurilor (Înțelegere verbală, Organizare perceptivă, Memorie de lucru și Viteza procesării) care pot servi la organizarea rezultatelor obținute la 11 dintre cele 14 subteste din WAIS — III în domenii separate ale funcționării cognitive.

În contrast cu tipul de AFC realizată cu datele WAIS — III, alte studii AFC au presupus o activitate mai elementară, care urmărea clarificarea organizării trăsăturilor cognitive și de personalitate. Un exemplu de astfel de studiu poate fi găsit în descrierea

pe care Gustafsson (2002) o face reanalizării datelor colectate de Holzinger și Swineford în 1930 de la un grup de elevi din clasele a șaptea și a opta (N= 301) testați cu o baterie alcătuită din 24 de teste concepute pentru a investiga aptitudinile din cinci arii largi (verbală, spațială, memorie, rapiditate și deducție matematică). Utilizând două modele diferite ale AFC, Gustafsson a obținut rezultate care susțin suprapunerea presupusă între factorii *G* (inteligență generală) și *Gf* (inteligență fluidă sau aptitudine de raționament) sugerați de analizele realizate în anii '30. De asemenea, Gustafsson a utilizat contrastele prezente în tiparele rezultatelor analizei factoriale inițiale, realizată în 1930, și AFC contemporană realizată de el pentru a ilustra implicațiile semnificative pe care diferite abordări ale instrumentelor, ca și componența eșantionului, le au pentru validarea constructului scorurilor obținute la testele de aptitudini. Analiza factorială confirmatorie și alte tehnici de modelare structurală se află încă în dezvoltare și sunt departe de a oferi o concluzie definitivă. Cu toate acestea, combinarea modelelor teoretice, a observațiilor empirice și a analizelor statistice sofisticate caracteristice acestor tehnici oferă mari promisiuni de avansare a înțelegerii instrumentelor pe care le utilizăm sau a relației dintre scorurile unui test și constructele pe care acesta este destinat să le evalueze.

Dovezi ale validității bazate pe relațiile dintre scorurile unui test și criteriu

Dacă obiectivele testării psihologice s-ar limita pur și simplu la descrierea performanței subiecților testați în termenii cadrelor de referință discutate în Capitolul 3 sau la sporirea înțelegerii constructelor psihologice și a relației lor reciproce, sursele de

dovezi discutate până acum ar trebui să fie suficiente. Totuși, interpretarea validă scorurilor unui test presupune deseori ca o semnificație inerentă scorurilor — fie că se bazează pe norme, pe conținutul testului, procese de răspuns, tipare determinate de convergență și divergență sau orice altă combinație a acestor surse de dovezi — să fie utilizată în inferențele pragmatice necesare pentru a lua decizii cu privire la oameni. Într-un astfel de caz, dovezile validității trebuie să evalueze semnificația pe care scorurile unui test o pot avea pentru aspecte care depășesc acele scoruri sau în domenii care se află în afara ariei directe a unui test. Cu alte cuvinte, trebuie să se arate că scorurile la test corelează cu diferite criterii utilizate în luarea deciziilor și realizarea predicțiilor.

Câteva fapte esențiale despre criterii

Merriam-Webster's Collegiate Dictionary (1995) definește *criteriul* drept „un standard pe care se poate baza o judecată sau o decizie" sau „un semn sau o trăsătură caracteristică". Deși forma de plural, criterii, este deseori utilizată adecvat în sensul de singular, în contextul prezent este necesar să utilizăm ambele forme ale cuvântului, deoarece sunt centrale pentru scopurile noastre. Un criteriu poate fi definit, mai larg, și drept ceea ce dorim *cu adevărat* să știm. Această ultimă definiție, deși mai puțin formală decât cea din dicționar, subliniază contrastul dintre ceea ce ne spun scorurile testului și motivele practice pentru care utilizăm teste.

În cazul testelor psihologice utilizate pentru a emite aprecieri sau a lua decizii cu privire la oameni, dovada unei relații dintre scorurile unui test și dimensiuni ale criteriului constituie o bază indispensabilă pentru evaluarea validității, dar nu suficientă în mod necesar. *Dimensiunile criteriului* sunt indicatori ai criteriilor

pe care testele sunt desemnate să le evalueze sau să le prezică și care sunt acumulate independent de testul în chestiune. Info 5.7 oferă o listă a tipurilor de criterii utilizate în mod tipic în validarea scorurilor unui test. Din moment ce natura criteriilor depinde de întrebările la care se dorește să se răspundă cu ajutorul testelor, rezultă că procedurile de validare bazate, parțial sau în întregime, pe relațiile dintre scorurile unui test și dimensiunile criteriului

Info 5.7

Criterii tipice utilizate în validarea scorurilor unui test

Deși există un număr aproape infinit de dimensiuni ale criteriului care pot fi utilizate în validarea scorurilor unui test, în funcție de scopurile testării, cele mai frecvente categorii sunt următoarele:

- *Indicatori ai realizărilor academice sau ai performanței într-o formă de instruire specializată*, cum ar fi calificative școlare, diplome de absolvire, distincții, premii sau demonstrații ale competenței în aria de instruire prin performanță ridicată (e.x.: cântat la pian, activitate mecanică, zbor, programare, examen de admitere în profesia de avocat și teste de certificare)
- *Indicatori ai performanței în muncă* cum ar fi înregistrări de vânzări, înregistrări ale producției, promovări, creșteri de salariu, longevitate în posturi care cer competență, performanță lipsită de accidente la locul de muncă sau evaluări realizate de supervizori, colegi, studenți, clienți și așa mai departe.
- *Apartenența la grupuri contrastante* bazate pe diagnostic psihiatric, statut ocupațional, achiziții eduacaționale și orice altă variabilă relevantă.
- *Evaluări* ale comportamentului sau *trăsăturilor* de personalitate realizate de observatori independenți, rude, colegi sau alți asociați care au baze suficiente pentru a le oferi.
- *Scoruri obținute la alte teste relevante.*

trebuie să producă dovezi ale unei legături dintre *predictori* (scorurile la test) și criterii.

Dimensiunile sau estimările criteriului pot fi *dihotomice* în mod natural (e.g. absolvire vs. renunțare prematură) sau pot fi *dihotomizate* artificial (succes versus eșec), *politomice* (diagnostic de tulburare de anxietate versus tulburare de dispoziție vs. tulburare disociativă, apartenența la ocupații artistice vs. științifice vs. literare); sau *continue* (punctajul mediu, numărul de unități vândute, scoruri la un inventar de depresie etc.). În timp ce natura criteriilor depinde de deciziile sau predicțiile care urmează a fi făcute cu ajutorul scorurilor testului, metodele utilizate pentru a stabili relații între scorurile testului și criterii variază în funcție de caracteristicile formale ale scorurilor testului și ale dimensiunilor criteriului. În general, atunci când dimensiunea criteriului este exprimată în manieră dihotomică (succes vs. eșec) sau în termenii unui sistem de categorii (apartenența la unul dintre grupurile contrastante), validitatea scorurilor unui test este evaluată în termenii *ratei succesului*. În mod tipic, rata succesului indică procentul de decizii sau de clasificări corecte realizate utilizând scorurile unui test, deși pot fi utilizate și diferențele medii și indicatori potriviți ai corelației. Atunci când dimensiunile criteriului sunt continue (scoruri la testele de achiziții, calificative, evaluări etc.) instrumentele principale utilizate pentru a indica mărimea relației dintre scorurile unui test și dimensiunea criteriului sunt coeficienții de corelație. Cu toate acestea, dacă o anumită valoare a unui criteriu continuu, cum ar fi o notă medie 5,0, este utilizată ca limită pentru a determina un rezultat specific, cum ar fi absolvirea unui colegiu, scorurile la un test utilizat ca predictor pot fi evaluate în termenii diferențierii între cei care

îndeplinesc sau depășesc limita impusă de criteriu și cei care nu o fac.

Istoria testării psihologice din ultimele decenii reflectă nu numai o evoluție a înțelegerii naturii și limitelor testelor și scorurilor obținute la teste, dar și o apreciere sporită a semnificației și complexității dimensiunilor criteriului (vezi James, 1973; Tenopyr, 1986; Wallace, 1965). Ca urmare, cu rare excepții, ideea că există „un criteriu" în raport cu care un test poate fi validat nu se susține mai mult decât propoziția că validitatea unui test poate fi determinată pe o bază de tipul tot — sau — nimic. În prezent sunt general înțelese următoarele fapte cu privire la criterii:

1. În cele mai multe studii de validare, există mulți indicatori posibili (atât cantitativi, cât și calitativi) care pot fi calificați drept dimensiuni ale criteriului, inclusiv scoruri obținute la alte teste decât cele validate. Prin urmare, trebuie să se acorde o mare atenție selecției criteriilor și măsurilor criteriilor.
2. Unele dimensiuni ale criteriilor sunt mai fidele și mai valide decât altele. Astfel, fidelitatea și validitatea dimensiunilor criteriilor trebuie să fie evaluate, așa cum se întâmplă și cu fidelitatea și validitatea scorurilor unui test.
3. Unele criterii sunt mai complexe decât altele. Ca rezultat, poate exista sau nu o corelație între dimensiunile criteriilor, mai ales atunci când criteriile sunt multifațetate.
4. Unele criterii pot fi determinate în momentul testării; altele apar în timp. Rezultă că pot exista sau nu corelații substanțiale între dimensiunile criteriului disponibile la scurt timp după testare și criterii mai distanțate în timp, care pot fi evaluate doar după o perioadă mai lungă.

5. Relațiile dintre scorurile unui test și dimensiunile unui criteriu pot sau nu să fie generalizate la grupuri, cadre sau perioade de timp. Prin urmare, dovezile validității în raport cu criteriul trebuie să fie demonstrate din nou pentru populații care diferă de eșantioanele inițiale de validare în moduri care influențează relația dintre scorurile obținute la un test și criterii, precum și pentru diferite cadre și momente.
6. Forța sau calitatea dovezilor validității cu privire la evaluarea sau predicția unui criteriu depinde atât de caracteristicile testului, cât și de caracteristicile dimensiunilor criteriului. Dacă dimensiunile criteriului nu sunt fidele sau sunt arbitrare, indicatorii validității scorurilor testului vor fi slabi, indiferent de calitatea testului utilizat pentru a evalua sau a face predicții în raport cu criteriul.

Procedurile de validare în raport cu criteriul

Deciziile în raport cu criteriul utilizează scorurile obținute la test pot fi clasificate în două tipuri de bază: (a) cele care presupun determinarea statutului curent al unei persoane și (b) cele care presupun predicția performanțelor sau comportamentelor viitoare. Într-un anumit sens, această dihotomie este artificială pentru că, indiferent dacă trebuie să aflăm ceva despre statutul curent sau despre performanța viitoare a unei persoane, singura informație pe care scorurile unui test o pot transmite derivă din comportamentul prezent — adică din performanța subiectului în momentul testării. Cu toate acestea, procedurile de validare în raport cu criteriul sunt clasificate frecvent drept concurente sau predictive, în funcție de dimensiunile utilizate și de obiectivele lor principale.

Validarea concurentă și validarea predictivă

Dovezi pentru *validarea concurentă* sunt culese atunci când indicatorii criteriilor pe care scorurile testelor sunt destinate să le măsoare sunt disponibili în momentul în care sunt realizate studiile de validare. Strict vorbind, validarea concurentă este adecvată pentru scorurile unui test care va fi utilizat în determinarea statutului curent al unei persoane din cadrul unei anumite scheme de clasificare, cum ar fi categoriile de diagnostic sau nivelele de performanță. Pe de altă parte, dovezile *validării predictive* sunt relevante pentru scorurile unui test care urmează să fie folosit în luarea deciziilor pe baza estimării nivelelor de performanță sau a rezultatelor comportamentale viitoare. În mod ideal, procedurile de validare predictivă necesită colectarea datelor la o variabilă predictor (scorurile la test) și se așteaptă ca datele pentru criteriu să devină disponibile astfel încât cele două seturi de date să poată fi corelate. Acest proces este deseori nepractic datorită elementului timp implicat în așteptarea „maturizării" criteriului și, de asemenea, datorită dificultății de a găsi eșantioane adecvate utilizării în astfel de studii. Ca rezultat, validarea concurentă este deseori utilizată ca substitut al validării predictive, chiar și în cazul testelor care vor fi utilizate pentru a estima performanțe viitoare, cum ar fi admiterea la colegiu sau testele anterioare angajării. În aceste cazuri, testul aflat în construcție este administrat unui grup de persoane, cum ar fi studenți în al doilea an sau angajați, pentru care datele criteriului sunt deja disponibile.

Deseori distincția dintre cele două tipuri de proceduri de validare pivotează în jurul modului în care utilizatorii testului pun întrebările la care doresc să răspundă cu ajutorul unui test. Info 5.8 conține exemple de câteva întrebări tipice și situații de luare a deciziilor care pot cere fie dovezi ale validității predictive, fie ale

Info 5.8

Relația dintre întrebări, decizii și predicții care necesită validare în raport cu criteriul

Întrebări legate de statutul curent	Obiectivul imediat: decizii	Întrebări implicite: predicții
John X. suferă de schizofrenie, tulburare de panică, depresie clinică, deficit de atenție sau altă tulburare psihică?	John X. ar trebui să primească un tratament (medicamente, psihoterapie etc.) recomandat pentru tulburarea în chestiune?	John X. va profita (puțin, mult sau deloc) în urma tratamentului recomandat?
Mary Y. poate simți impulsuri suicidare (sau omucidare) pe care să nu fie capabilă să le depășească dacă este lăsată să facă ce dorește fără a fi supravegheată?	Ar trebui ca Mary să continue să fie spitalizată (sau închisă)?	Mary va încerca să se sinucidă (sau să ucidă o altă persoană) dacă este lăsată în libertate?
Joe Z. este supradotat (sau suferă de retard mental grav)?	Joe Z. ar trebui admis într-un program de educație specială pentru indivizi supradotați (sau cu retard mental)?	Joe Z. va beneficia (puțin, mult, deloc), datorită programului, de educație specială?
Tom P. este onest și de încredere (sau motivat pentru a lucra în domeniul vânzărilor)?	Tom P. ar trebui să fie angajat pentru a lucra pe post de casier (sau vânzător)?	Tom P. va fi un casier conștiincios (sau un vânzător cu succes) sau va fura bani (sau va face vânzări reduse)?
Jane Q. este capabilă să desfășoare o activitate de nivelul colegiului (sau să piloteze un avion)?	Jane Q. ar trebui să fie admisă la colegiu (sau să i se acorde un brevet de pilot)?	Jane Q. va fi capabilă să termine colegiul cu o notă medie suficient de mare pentru a absolvi (sau va fi capabilă să piloteze un avion fără a provoca accidente)?

validității concurente, în funcție de modul în care este pusă întrebarea și de orizontul de timp ales. Pentru a ilustra diferența dintre strategiile de validare concurentă și predictivă, va fi prezentat câte un exemplu relativ simplu pentru fiecare tip de studiu, urmat de o discuție a aspectelor majore relevante pentru validarea relativă la criteriu.

Exemplu de validare concurentă: Whitaker Index of Schizophrenic Thinking

Testele utilizate pentru identificarea persoanelor cu tulburări psihice grave, cum ar fi schizofrenia sau depresia, de obicei trec printr-un proces de validare concurentă. În mod tipic, aceste studii utilizează două sau mai multe eșantioane de indivizi care diferă în privința statutului diagnostic stabilit în mod independent. Unul dintre numeroasele instrumente ale căror scoruri sunt validate în această manieră este Whitaker Index of Schizophrenic Thinking (WIST; Whitaker, 1980). WIST a fost construit pentru a identifica tipul de deteriorare a gândirii care acompaniază deseori

sindroamele schizofrenice. Fiecare dintre cele două forme ale sale (A și B) constă din 25 de itemi cu alegere multiplă.

În standardizarea WIST, Whitaker a utilizat eșantioane alcătuite din pacienți cu schizofrenie acută și cronică (S) și trei grupuri de persoane fără schizofrenie (NS) pentru a obține scorurile-prag care ar diferenția optim indivizii S de cei NS. Scorurile-prag stabilite cu ajutorul grupurilor de standardizare discriminează între grupurile S și NS cu o eficiență de 80 % pentru forma A și 76% pentru forma B. Astfel, în funcție de formă, indicatorul de prag a dus la decizii incorecte cuprinse între 20% și 24%. În cazul formei A, deciziile incorecte *fals negative* — cele în care subiecții S au fost clasificați ca NS de către indicator — au fost mult mai multe (33%) decât deciziile *fals pozitive*, în care subiecții NS au fost clasificați ca S (10%). Același tipar (30% fals negative vs. 13% fals pozitive) a fost obținut și în cazul formei B. Vezi capitolul 7 pentru alte explicații cu privire la terminologia utilizată pentru a denumi deciziile incorecte (decizii fals pozitive și fals negative).

Dovezile suplimentare ale validării pentru WIST au inclus studiile realizate în Mexic și Spania cu traduceri în limba spaniolă ale instrumentului. Aceste studii au indicat o anumită susținere a capacității WIST de a detecta deteriorările gândirii asociate cu schizofrenia, în ciuda ratelor diferite ale eficienței și a scorurilor-prag diferite.

Rata de succes obținută când scorurile prag WIST sunt utilizate pentru a diferenția subiecții S de cei NS este destul de tipică pentru testele de acest tip. Datorită ratelor de eroare relativ înalte pe care le pot atinge scorurile-prag WIST, ele nu ar trebui niciodată utilizate ca unic vehicul în stabilirea unui diagnostic de schizofrenie, ca în cazul oricărui alt indicator unic. Cu toate acestea, în funcție de cadrul și de contextul testării, WIST se poate dovedi util ca parte a unei baterii de triere sau ca indicator al schimbării în simptomatologia pacienților diagnosticați cu schizofrenie. Deși discriminările realizate cu ajutorul scorurilor-prag WIST sunt departe de a fi perfecte, ele pot fi utile în investigarea posibilității ca o persoană să sufere de o deteriorare a gândirii. Pentru o trecere în revistă a WIST, vezi Flanagan (1992).

ATENȚIE!

- Nu este de așteptat ca rezultatele testelor psihologice, ca și cele ale testelor medicale și ale testelor din multe alte domenii să aibă o precizie de 100%.
- Din cauza naturii mai puțin decât perfecte a tuturor indicatorilor de diagnostic, psihologii angajați în evaluarea clinică — ca și diagnosticienii din alte domenii — nu ar trebui să se bazeze niciodată pe un singur indicator.

Validarea predictivă: un exemplu ipotetic

Testele utilizate pentru predicția performanțelor în cadre ocupaționale și educaționale necesită strategii de validare adaptate predicției. Modelul ideal al unui studiu de validare predictivă în oricare dintre aceste domenii va presupune următorii pași: (a) testarea unui grup neselectat de candidați cu un test sau o baterie de teste de capacitate, (b) angajarea sau admiterea lor fără a lua în considerare scorurile obținute la test, (c) o perioadă de așteptare până când dimensiunile criteriului de performanță în muncă sau performanță academică ajung să fie disponibile, (d) obținerea corelațiilor între scorurile la testele de admitere sau pre-angajare și dimensiunile criteriului și (e) utilizarea datelor corelaționale pentru a obține o ecuație de regresie care să estimeze sau să facă predicții cu privire la performanța la criteriu a viitorilor candidați. Din motive evidente, cei mai mulți angajatori sau administratori de școli nu sunt dispuși să angajeze sau să accepte

toți candidații — mai ales atunci când numărul acestora depășește numărul de locuri disponibile — pentru a realiza un studiu de validare. Cu toate acestea, pentru simplitate și pentru a ilustra modul în care pot fi făcute predicții pe baza scorurilor la un test, va fi descris un exemplu ipotetic de astfel de studiu.

Pentru a oferi acestui exemplu un anumit context, să spunem că — pentru a asigura un nivel profitabil de producție — proprietarul unei mici fabrici dorește să angajeze oameni care sunt capabili să proceseze cel puțin 50 de dispozitive pe oră la o linie de asamblare. Deoarece nu există alte cerințe pentru această muncă, pe lângă punctualitatea la începutul programului, criteriul în raport cu care trebuie realizată predicția este pur și simplu numărul de dispozitive produse pe oră la linia de asamblare. Rezultatul la linia de asamblare depinde în principal de viteza și precizia lucrătorilor în cazul sarcinilor manuale. Prin urmare,

Tabelul 5.4 Datele pentru exemplul de validare predictivă

Candidat	Scorul la test (X)	Rezultat (Y)	$X-M_x$ (x)	$Y-M_y$ (y)	x^2	y^2	xy
1	18	56	5	6	25	36	30
2	12	50	-1	0	1	0	0
3	8	47	-5	-3	25	9	15
4	20	52	7	2	49	4	14
5	14	52	1	2	1	4	2
6	5	42	-8	-8	64	64	64
7	10	48	-3	-2	9	4	6
8	12	49	-1	-1	1	1	1
9	16	50	3	3	9	0	0
10	15	54	2	2	4	16	8
Suma	130	500	0	0	188	138	140
Media	13	50					

un test de dexteritate manuală este selectat drept predictor. Tabelul 5.4 indică *datele bivariate* pentru 10 candidați ipotetici pentru locul de muncă la linia de asamblare care au fost supuși unui *test de dexteritate manuală*, au fost angajați și instruiți și li s-a măsurat *rezultatul pe oră la linia de asamblare* în ocazii repetate și s-a calculat o medie a acestor rezultate pentru a avea o dimensiune viabilă a criteriului.

Așa cum am discutat în Capitolul 2, atunci când două variabile prezintă o relație liniară și o corelație puternică una cu cealaltă, este posibilă predicția uneia pe baza cunoașterii celeilalte prin aplicarea modelului regresiei liniare. Figura 5.1 prezintă diagrama

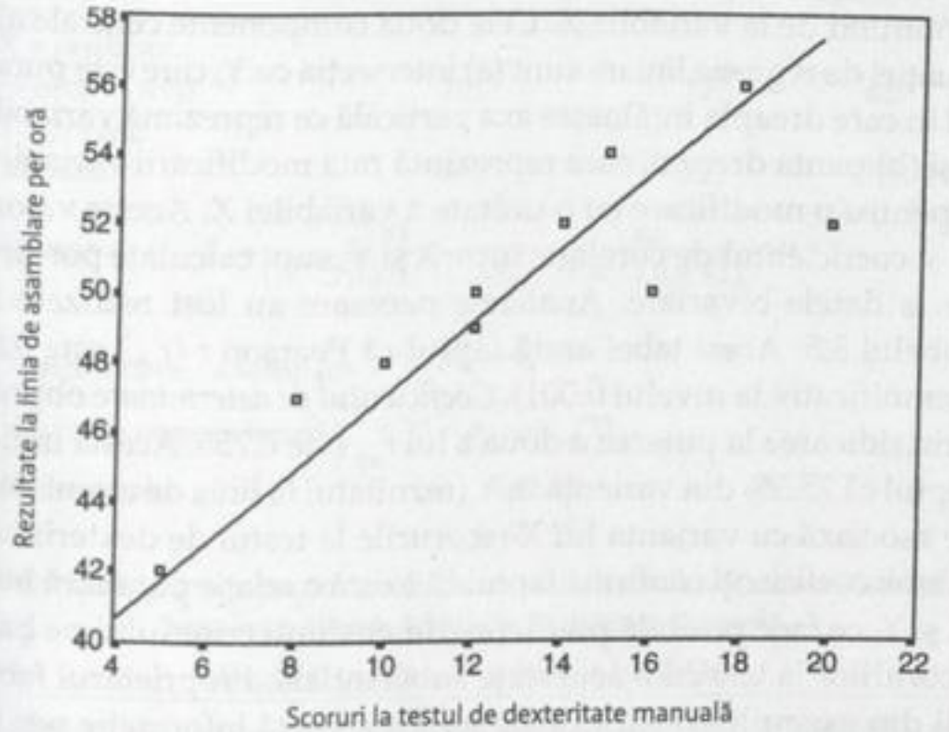

Figura 5.1. Diagrama de împrăștiere a datelor și dreapta de regresie pentru exemplul de validare predictivă

de împrăștiere a datelor bivariate din Tabelul 5.4 și indică faptul că relația dintre scorurile obținute la testul de dexteritate manuală și rezultatul la linia de asamblare per oră este puternică, pozitivă și liniară.

În acest caz, este necesar să rezolvăm o ecuație de regresie liniară pentru predicția criteriului rezultat la linia de asamblare (variabila Y) pornind de la scorurile obținute la testul de dexteritate manuală (variabila X). Această ecuație exprimă relația dintre X și Y și conține cele două componente majore necesare pentru a forma *dreapta de regresie*, ilustrată în Figura 5.1. Dreapta de regresie este dreapta care se potrivește cel mai bine datelor bivariate prin faptul că minimizează erorile predicției variabilei Y pornind de la variabila X. Cele două componente centrale ale ecuației de regresie liniare sunt (a) intersecția cu Y, care este punctul în care dreapta întâlnește axa verticală ce reprezintă variabila Y și (b) panta dreptei, care reprezintă rata modificării variabilei Y pentru o modificare cu o unitate a variabilei X. Aceste valori, ca și coeficientul de corelație între X și Y, sunt calculate pornind de la datele bivariate. Analizele necesare au fost realizate în Tabelul 5.5. Acest tabel arată faptul că Pearson *r* (r_{xy}) este 0,87 (semnificativ la nivelul 0,001). *Coeficientul de determinare* obținut prin ridicarea la puterea a doua a lui r_{xy} este 0,755. Acesta indică faptul că 75,5% din varianța în Y (rezultatul la linia de asamblare) se asociază cu varianța lui X (scorurile la testul de dexteritate). Ambii coeficienți confirmă faptul că există o relație puternică între X și Y, ce face posibilă predicția dimensiunii criteriului pe baza scorurilor la test cu o acuratețe substanțială. Proprietarul fabricii din exemplul nostru poate aplica această informație pentru a estima rezultatele la linia de asamblare ale viitorilor candidați prin administrarea testului de dexteritate manuală și inserarea

scorurilor obținute de aceștia în ecuația de regresie după cum urmează.

Să presupunem că a 11-a persoană care a candidat pentru un loc de muncă (după cei 10 care au fost utilizați în studiul ipotetic de validare) obține la testul de dexteritate manuală scorul 17. Utilizând coeficienții prezentați în Tabelul 5.5 pentru a rezolva ecuația de regresie $Y' = 40{,}32 + 0{,}745\,(17)$, proprietarul fabricii va afla că rezultatul estimat la linia de asamblare per oră pentru cel

Tabelul 5.5 Analiza datelor validării predictive

Statistica descriptivă	Numărul de observații (*N*)	Medie	Abatere standard (SD)
X = predictor (scoruri la test)	10	13	4,57
Y = criteriu (rezultat)	10	50	3,91
Pearson *r*	$r_{xy} = \frac{\Sigma xy}{(N-1)(SD_x)(SD_y)} = \frac{140}{(9)(4{,}57)(3{,}91)} = 0{,}87$		
Coeficientul de determinare	$r^2_{xy} = 0{,}755$		
Ecuația de regresie liniară[a]	$Y' = a_{yx} + b_{yx}(X)$		
Date exemplu	$a_{yx} = 40{,}32$ $b_{yx} = 0{,}745$		

Ecuația de regresie pentru predicția rezultatelor la linia de asamblare pe baza scorurilor la testul de dexteritate manuală

Rezultat prezis = $Y' = 40{,}32 + 0{,}745\,(X)$

[a] Unde Y' = scorul prezis al criteriului; $a_{yx} = M_y - b_{yx}(M_x)$ = intersecția cu dreapta de regresie; $b_{xy} = (\Sigma_{xy}) / (\Sigma_{x^2})$ = panta dreptei de regresie; și X = scorul la predictor (scorul la testul de dexteritate manuală).

de-al 11-lea candidat este de 53 de dispozitive, cu trei mai mult decât numărul minim dorit.

Corelațiile dintre scorurile obținute la test și dimensiunile criteriului (r_{xy}) sunt numite de obicei *coeficienți de validitate*. Dacă r_{xy} ar fi 1,00, indicând o corelație perfectă între scorurile obținute la test și dimensiuniel criteriului, ar fi posibilă o predicție perfectă. Deși $r_{xy} = 0,87$ obținut cu datele din exemplul nostru este extrem de înalt — mult mai înalt decât coeficienții de validitate obținuți în mod tipic cu date reale — el nu este perfect. Astfel, există o anumită eroare în estimările criteriului realizate utilizând scoruri obținute la test. Această eroare este stabilită prin eroarea standard a estimării (SE_{est}), o mărime statistică ce exprimă, pe scala utilizată pentru măsurarea criteriului, eroarea predicțiilor bazate pe corelații imperfecte.

Formulele pentru SE_{est} și pentru un termen de corecție aplicat lui SE_{est} datorită mărimii reduse a eșantionului sunt prezentate în Info 5.9 alături de rezultatele pentru datele din exemplul nostru. Interpretarea lui SE_{est} presupune (a) că scorul prezis al criteriului, sau Y', este valoarea medie a unei distribuții ipotetice normale a tuturor scorurilor posibile ale criteriului pentru candidatul în chestiune și (b) că SE_{est} este abaterea standard a acelei distribuții. Aceste premize ne permit să atașăm un nivel de probabilitate predicțiilor făcute pornind de la scorurile obținute la test pe baza ariilor de sub curba normală din Anexa C. Pentru aceasta, se calculează intervalul cuprins între $Y' \pm SE_{est}(z)$, unde z este valoarea care corespunde nivelului dorit de probabilitate. Pentru cel de-al 11-lea candidat din exemplul nostru, scorul 17 la testul de dexteritate a dus la predicția unui Y' de 53. SE_{est} pentru scorurile la test, calculată în Info 5.9. este 2,05 (sau 2). Prin urmare, putem face predicția că șansele ca acest candidat să producă între

Nu uitați!

Evaluarea adecvării coeficienților de validitate de diferite mărimi este o problemă relativă care depinde de scopul pentru care vor fi utilizate scorurile obținute la test. În general, există două moduri majore de abordare a acestei probleme:

1. Atunci când validitatea este exprimată sub forma unui coeficient de corelație (r_{xy}), proporția de varianță comună pentru predictor (X) și criteriu (Y) este deseori estimată prin ridicarea lui (r_{xy}) la puterea a doua și obținerea unui *coeficient de determinare* sau r^2_{xy}. Coeficientul de determinare, discutat în Capitolul 2, exprimă proporția de varianță a lui Y asociată cu varianța lui X. De exemplu, coeficientul de validitate predictivă de 0,87 obținut în exemplul prezentat în Tabelul 5.5 indică faptul că scorurile la testul de dexteritate manuală utilizate ca predictor pot explica aproximativ 75% (0,87 x 0,87 = 0,76) din varianța criteriului rezultate obținute la linia de asamblare.
2. Coeficienții de validitate pot fi evaluați și în termenii *validității incrementale* aduse de un test — adică măsura în care utilizarea unui test (sau a oricărui alt instrument) duce la sporirea eficienței deciziilor luate într-o situație dată comparativ cu validitatea altor metode. Acest aspect al validității privește utilitatea unui test și este discutat în Capitolul 7.

Coeficienți cu valoarea de 0,20 sau 0,30 nu sunt neobișnuiți în studiile de validitate predictivă. Ca regulă generală, coeficienții cu valoarea de 0,40 sau mai mari sunt considerați acceptabili. Deși unii dintre ei pot părea scăzuți, ei trebuie priviți în contextul naturii multideterminate a performanței în cele mai multe medii *și* trebuie cântăriți în lumina eficienței predictive a celorlalte metode de luare a deciziilor de selecție.

51 și 55 de dispozitive pe oră (53 ± 2) la linia de asamblare sunt 68/100. Pentru a face o predicție la un nivel de încredere de 95% (în loc de 68%) ar trebui să utilizăm o valoare z de 1,96, care

Info 5.9

Eroarea standard a estimării (SE_{est})

$$SE_{est} = SD_y \sqrt{1 - r_{xy}^2}$$

SE_{est} pentru exemplul de validare predictivă:

$$SE_{est} = 3{,}91 \sqrt{1 - 0{,}755} = 3{,}91\,(0{,}495) = 1{,}935$$

SE_{est} corectată pentru mărimea eșantionului:

$$SE_{est} \text{ corectată} = SE_{est} \sqrt{\frac{N-1}{N-2}} = 1{,}935 \sqrt{\frac{10-1}{10-2}} = 1{,}935\,(1{,}06) = 2{,}05$$

demarchează nivelul de semnificație de 0,05 pentru un test bilateral. Rezolvând ecuația $Y' \pm SE_{est}$ (1.96), putem face predicția că există 95/100 șanse ca acest candidat să producă între 49 și 57 de dispozitive pe oră (53 ± 4).

Aspecte privind studiile de validare în raport cu criteriul

Unele dintre dificultățile inerente întâlnite în studiile de validare în raport cu criteriul au fost menționate anterior în acest capitol odată cu prezentarea noțiunii de criteriu. În continuare sunt prezentate considerații suplimentare care merită să fie examinate complet de către potențialii utilizatori ai instrumentelor bazate pe acest tip de dovezi ale validității, utilizând ca punct de plecare unele dintre caracteristicile exemplelor descrise anterior. Deși scopurile noastre nu ne permit o explorare extensivă a acestor aspecte — sau a diferitelor metode care pot fi utilizate în aceste cazuri — unele dintre cele mai importante trebuie descrise pe scurt.

Caracteristici ale dimensiunilor criteriului

Așa cum s-a menționat anterior, criteriile în funcție de care sunt validate scorurile unui test pot diferi mult în termenii propriilor fidelități și validități. În cazul WIST, criteriul utilizat pentru stabilirea validității scorurilor a fost statutul diagnostic (S vs. NS) al indivizilor din eșantioanele de validare. În cazul în care clasificarea inițială a subiecților care au participat la aceste studii a inclus anumite diagnostice incorecte (unii S care erau NS sau invers), datele validității ar fi, în mod evident, mai slabe. O precauție similară se aplică tuturor procedurilor de validare care se bazează pe criterii subiective, cum ar fi evaluările sau alte judecăți calitative utilizate pentru a include oamenii în categorii ale grupurilor-criteriu. În plus, validitatea dimensiunilor criteriului poate fi slăbită atunci când cei însărcinați cu determinarea poziției în raport cu criteriul a indivizilor din eșantioanele de validare au acces la scorurile obținute la testul utilizat ca predictor. Acest tip de eroare, cunoscută drept *contaminarea criteriului*, poate fi prevenită ușor asigurându-ne că profesorii, supervizorii, diagnosticienii și persoanele care fac evaluări sau emit judecăți legate de dimensiunile criteriului nu cunosc scorurile obținute la test și nu sunt influențați acestea. Cu privire la evaluări, ca și în cazul altor dimensiuni ale criteriului care depind de judecata subiectivă, autorii testului trebuie să ofere dovada că instrumentele și metodele utilizate pentru a cota sau a clasifica grupurile-criteriu din studiile de validare sunt, la rândul lor, fidele și valide. Atunci când criteriul constă din apartenența la un grup cum ar fi o anumită categorie de diagnostic, fidelitatea și validitatea sa poate fi îmbunătățită prin selectarea atentă a subiecților pe baza dovezilor provenite din surse multiple și, preferabil, independente. În ceea ce privește cotarea criteriilor, trebuie asigurată

fidelitatea și validitatea acestora. În acest scop există o literatură extinsă dedicată modului de instruire a evaluatorilor astfel încât să se minimizeze influența prejudecăților și explorării celor mai bune formate și metodologii de cotare (vezi Guion, 1998, cap. 12).

Studiul ipotetic al validității scorurilor testului de dexteritate manuală — menit să ilustreze într-un mod cât mai simplu posibil procedurile de validare predictivă — conține câteva trăsături nerealiste. De exemplu, criteriul „rezultate la linia de asamblare" poate fi stabilit fidel și precis prin numărarea dispozitivelor produse de lucrători. Nu toate criteriile sunt atât de simplu sau ușor de stabilit. Succesul în multe medii ocupaționale, cum ar fi management, practica medicală, învățământ și așa mai departe poate fi judecat pe baza unor criterii care diferă în termenii fidelității măsurării lor, ai gradului de control pe care un lucrător îl are în raport cu ele și al valorii pe care organizația le-o acordă. De exemplu, succesul unui manager poate fi determinat, printre altele, în termenii (a) productivității angajatului și (b) satisfacției angajatului. Abilitățile și caracteristicile personale care fac un manager să aibă succes în raport cu acestea (a) nu sunt în mod necesar identice cu cele care conduc la succes și (b) în anumite situații aceste două criterii pot intra chiar în conflict unul cu celălalt. Mai mult, fiecare dintre aceste criterii poate fi evaluat prin diferite metode. Productivitatea poate fi determinată de cantitatea sau de calitatea producției sau de ambele; satisfacția angajaților poate fi măsurată prin evaluările supervizorilor, prin rata de schimbare a personalului dintr-o unitate și așa mai departe. Scorurile obținute la test care fac predicții pentru o fațetă a criteriului pot să nu coreleze sau chiar să coreleze negativ cu cele care fac predicții pentru o altă fațetă.

Utilizarea predictorilor multipli

Abordarea tradițională a predicțiilor criteriilor complexe, precum performanța în muncă, a fost utilizarea unei baterii de teste. În acest context, termenul *baterie* se referă la o combinație de predictori selectați în mod special pentru predicția unuia sau mai multor criterii. Acest sens contrastează cu utilizarea termenului în cadre clinice sau de consiliere, unde, de obicei, o baterie trimite la orice grup de teste care sunt administrate unui individ în procesul evaluării psihologice. Scorurile la predictorii separați dintr-o baterie de testare pot fi combinate în diferite moduri, în funcție de cerințele problemelor de selecție sau de clasificare. *Tehnicile regresiei multiple*, de exemplu, combină scorurile obținute la fiecare test din baterie, inserându-le într-o ecuație de regresie liniară care include o ponderare numerică pentru fiecare scor obținut la un test din baterie. *Ecuațiile de regresie multiplă* sunt extinderi ale metodei regresiei liniare simple prezentată în Tabelul 5.5, dar care presupun predictori multipli, nu unici. Ponderea pentru fiecare predictor este direct proporțională cu corelația sa cu criteriul și invers proporțională cu corelația sa cu ceilalți predictori din baterie, astfel încât scorurile testului care au cea mai mare validitate și cel mai mic grad de suprapunere cu alte scoruri au ponderea cea mai mare. Poate fi calculat un coeficient de corelație multiplă (R) care reprezintă corelația dintre combinația optim ponderată a scorurilor testului și criteriu. O procedură alternativă de combinare a scorurilor dintr-o baterie de teste este cea prin *analiza profilului*. Această metodă presupune stabilirea unui scor-prag pentru fiecare predictor — pe baza relațiilor acestuia cu criteriul — și duce la respingerea tuturor candidaților ale căror scoruri se situează sub pragul minim la fiecare test sau numai la testele care evaluează abilități considerate importante pentru obținerea unei performanțe la criteriu.

Fiecare dintre cele două metode descrise are anumite dezavantaje. Ecuațiile de regresie multiplă permit ca deficiențele la unul sau mai mulți predictori să fie compensate printr-o performanță superioară la alți predictori. Ponderarea specifică a predictorilor din aceste ecuații trebuie, de asemenea, verificată pe eșantioane independente de cele utilizate pentru a obține ecuațiile, pentru a vedea dacă se menține corelația multiplă (R). Repetarea relației predictor-criteriu în eșantioane separate — un proces cunoscut drept *validare încrucișată* — este necesară deoarece orice coeficient de corelație, indiferent de mărimea sa, este într-o anumită măsură dependent de o eroare specifică eșantionului. O anumită reducere a mărimii coeficientului R inițial, sau *contracție*, este de așteptat în cazul validării încrucișate. În măsura în care contracția este neglijabilă, ponderea inițială poate fi considerată suficient de stabilă pentru a fi aplicată fără alte investigații.

Unul dintre dezavantajele majore ale utilizării metodei de analiză a profilului alături de scorurile-prag este faptul că, de obicei, această metodă nu reușește să ia în considerare posibila lipsă de fidelitate a scorurilor (vezi Cuantificarea erorii scorurilor la test: Eroarea standard de măsurare, în Capitolul 4). O altă dificultate apare din faptul că existența scorurilor-prag multiple poate conduce la respingerea prea multor candidați, mai ales a acelor candidați din medii defavorizate care ar putea obține scoruri sub prag la unul sau mai multe teste de aptitudini, dar care ar putea fi capabili să depășească aceste deficiențe prin instruire sau datorită puternicei motivații de a reuși. În general, utilizarea scorurilor-prag este justificată doar în situații în care un deficit într-o abilitate specifică ar putea avea consecințe serioase și grave pentru performanța în muncă. O soluție posibilă, atunci când este cazul, este selecția pe baza scorurilor-prag doar în cazul

testelor care evaluează abilități de importanță deosebită pentru activitatea ce urmează a fi desfășurată și utilizarea ecuației de regresie pentru ceilalți predictori din baterie.

Problema intervalului restrâns din eșantioanele de validare

Așa cum am menționat mai sus, o altă caracteristică nerealistă a studiului ipotetic de validare prezentat la începutul acestui capitol este aceea că presupune un eșantion eterogen alcătuit din 10 candidați pentru care nu există dimensiuni ale criteriului în momentul testării. Cele mai multe studii de validitate predictivă nu se desfășoară în această manieră. Ele utilizează eșantioane alcătuite din indivizi pentru care datele criteriului sunt deja disponibile, cum ar fi angajați sau studenți care au ocupat deja postul sau au început programul de învățământ pentru care va fi folosit testul ce urmează a fi validat. Cu alte cuvinte, cea mai mare parte a studiilor utilizează strategii de validare concurentă pentru a obține dovezi ale validității predictive. Tipul de indivizi pentru care criteriul este deja disponibil diferă de cei pentru care, în cele din urmă, va fi folosit testul prin faptul că ei au fost deja selecționați pentru angajare sau au fost admiși într-un program de studiu *și* au rămas pe acel post sau în acea școală fără a fi concediați sau a renunța. Din acest motiv, aproape întotdeauna putem presupune că scorurile lor la testele predictor care se află în procesul validării și la criteriu se vor situa într-un interval mai restrâns decât cel care s-ar obține în cazul unui eșantion neselecționat de participanți. Ne putem aminti din Capitolul 2 că efectul unei restricții asupra intervalului unei variabile este reducerea mărimii coeficienților de corelație. Astfel, ca o consecință a restricționării intervalului, corelațiile dintre scorurile testului și criteriu

(i.e. coeficienți de validitate) care rezultă din aceste studii de validare retrospectivă sunt de obicei mai mici decât în cazul în care eșantioanele ar fi extrase dintr-o populație mai eterogenă, cum ar fi *toți* indivizii care candidează pentru posturile sau programele academice în chestiune.

Generalizarea validității

Mărimea indicatorilor validității predictive obținuți pentru scorurile unui test depinde de patru elemente de bază: (a) componența eșantioanelor de validare în termenii dimensiunii și variabilității; (b) natura și complexitatea criteriului pentru care se vor face predicții; (c) caracteristicile testului; și (d) interacțiunile dintre toate acestea. Din moment ce fiecare dintre acești patru factori poate modifica rezultatele studiilor de validare, utilizatorii testelor trebuie să le ia în considerare cu atenție înainte de a presupune că dovezile publicate ale validității scorurilor testului obținute dintr-un singur studiu sunt relevante pentru scopurile lor și pentru populațiile lor de subiecți.

Cu privire la componența eșantioanelor de validare, am discutat deja problemele legate de dimensiunile criteriului și restricțiile asupra intervalului. În mod frecvent, și o dimensiune redusă a eșantionului reprezintă o problemă. Cei mai mulți angajatori nu au un număr mare de angajați în aceeași categorie de posturi, iar rezultatele cercetărilor de validare bazate pe eșantioane reduse sunt mai predispuse la erori specifice eșantionului decât cele bazate pe eșantioane mari. Setul de date bivariate pentru cei 10 candidați din exemplul nostru ipotetic de validare predictivă a dus la un coeficient de validitate nerealist de mare de 0,87. Deși este posibil să se obțină corelații mari atunci când indivizii care

participă la studiile de validare nu sunt selecționați, iar criteriile sunt limitat formulate, ca în acest exemplu, realitatea este că studiile de validare locală realizate utilizând eșantioane mici, cu un interval limitat al scorurilor și criterii nefidele, duc în mod tipic la estimări scăzute și instabile ale corelației dintre predictor și criteriu.

Variabile moderatoare

Un aspect suplimentar legat de alcătuirea eșantioanelor din studiile de validare predictivă privește rolul posibil al variabilelor moderatoare. O *variabilă moderatoare* este orice caracteristică a unui subgrup de persoane dintr-un eșantion care influențează gradul de corelație dintre alte două variabile. Teoretic, aproape orice caracteristică demografică (sex, etnie, nivel de educație, clasă socială, localizare geografică etc.) sau trăsătură psihică (interese, motivație, nivel de anxietate etc.) poate acționa ca variabilă moderatoare în studiile de validitate predictivă și poate produce un efect de interacțiune care fie va scădea, fie va crește corelația predictor-criteriu. Pentru a investiga această posibilitate, este necesară desfășurarea unor studii de validare separate sau împărțirea eșantioanelor de validare în subgrupuri care să difere în raport cu variabila despre care se presupune că moderează coeficienții de validitate.

S-au descoperit consecvent diferențe considerabile în favoarea albilor și asiaticilor comparativ cu negrii sau hispanicii, diferențe ale scorurilor medii la testele de aptitudini academice obținute de persoane din diferite grupuri etnice sau rasiale. Aceste diferențe au trezit suspiciunea că rasa sau etnia poate modera validitatea predictivă a testelor de selecție. Ca rezultat, multe studii au realizat analize separate ale mărimii corelației predictor-criteriu

și ale coeficienților de regresie pentru albi, negri, hispanici și membri ai altor minorități rasiale și etnice. Scopul unor astfel de studii este de a stabili cu certitudine dacă scorurile unui test au validitate comparabilă pentru diferite grupuri și au valoare predictivă la fel de mare pentru toate sau dacă scorurile obținute la test au validități diferite pentru grupuri diferite și, prin urmare, sunt preferențiale. În acest context, termenul *preferențial* este utilizat pentru a indica orice diferență sistematică a relației dintre predictori și criterii în cazul persoanelor care aparțin unor grupuri diferite. Diferențele sistematice se pot manifesta în două moduri și anume validitate diferențială și predicție diferențială.

Validitatea diferențială, în contextul unei orientări preferențiale a testului, se referă la diferențele valorilor corelațiilor obținute între predictori și criterii pentru membrii unor grupuri diferite. Diferențele de mărime a coeficienților de validitate sugerează că scorurile testului duc la predicții mai precise pentru membrii grupului cu un coeficient mai mare. Dacă pantele dreptelor de regresie pentru cele două grupuri sunt diferite, avem dovada grafică a validității diferențiale; panta dreptei de regresie este mai mare pentru grupul cu un coeficient de validitate mai înalt. Din acest motiv, problema validității diferențiale este denumită și *preferință indicată de pantă* (vezi Tabelul 5.5 și Fig. 5.1).

Predicția diferențială, pe de altă parte, apare atunci când scorurile obținute la test duc la sub sau supra-predicții ale performanței la criteriu a unui grup comparativ cu altul. Problema este etichetată *preferință indicată de interceptare*, deoarece, atunci când un predictor supra-prezice sau sub-prezice performanța la criteriu pentru un grup, intersecția cu Y, sau punctul de origine al dreptei de regresie a acelui grup cu axa Y este diferit de cel al altor grupuri. Cu privire la problemele validității diferențiale și predicției

diferențiale a scorurilor unui test, sunt posibile câteva rezultate, deși nu la fel de probabile. Scorurile obținute la un test pot prezenta (a) nicio preferință în raport cu grupuri diferite, (b) validitate diferențială și predicție diferențială, (c) validitate diferențială fără predicție diferențială sau (d) predicție diferențială fără validitate diferențială. În general, căutarea dovezilor că rasa acționează ca variabilă moderatoare ce duce la validitate diferențială și predicție diferențială pentru membrii minorităților rasiale pe baza scorurilor la teste de capacitate nu a fost foarte fructuoasă. De fapt, studiile care au investigat diferențele dintre grupurile etnice în privința preciziei predicțiilor performanței la criteriu au indicat faptul că testele deseori tind la *supra-predicții* ale performanțelor negrilor și hispanicilor comparativ cu cele ale albilor și asiaticilor. Pe de altă parte, unele teste — mai ales cele utilizate în deciziile de admitere în domeniul educațional — uneori duc la sub-predicții ale performanțelor femeilor, deși într-o măsură mai mică decât supra-prezic performanța anumitor grupuri minoritare etnice sau rasiale (vezi Young, 2001; Zwick, 2002, cap. 5 și 6).

Bineînțeles, motivele pentru care anumite scoruri obținute la teste supra-prezic în general performanța la criteriu — tipic sub forma punctajului mediu — pentru membri ai anumitor grupuri minoritare etnice sau rasiale și deseori sub-prezic performanța femeilor au reprezentat subiectul multor presupuneri și dezbateri. Lipsa de fidelitate și orientarea preferențială care influențează criteriul exprimat prin note sunt citate în mod frecvent drept posibile explicații ale supra-predicției performanțelor academice ale minorităților etnice sau rasiale, așa cum se întâmplă și cu marile diferențe în educația primită în familie sau în calitatea experienței educaționale anterioare. O explicație nouă se centrează

în jurul noțiunii de *amenințare cu stereotipul*, care se referă la efectele dăunătoare pe care frica de confirmare a stereotipurilor negative radicale pare să le aibă asupra performanței la test a unor membri ai grupului minoritar (Steele, 1997). Cât despre sub-predicția performanței la colegiu a femeilor, cea mai frecvent citată presupunere ce centrează pe faptul că, în opoziție cu bărbații, femeile ca grup (a) tind să aleagă cursuri care sunt notate mai puțin exact sau (b) privesc cu mai multă seriozitate studiile lor. Deși s-ar putea să fie adevărat că aceste variabile și altele legate de statutul etnic sau de gen influențează scorurile obținute la test și performanța la criteriu a diferitelor grupuri, precum și corelațiile dintre acestea, nu este posibil întotdeauna să se determine cu precizie care sunt acești factori. Mai mult, se pare că măsura predicției diferențiale a performanței la colegiu pentru grupurile minoritare rasiale sau etnice a scăzut în ultimul sfert de secol (Young, 2001). În plus, indiferent care sunt aceste variabile, în mod evident ele nu sunt la fel de relevante pentru toți membrii acelor grupuri — care sunt destul de eterogene. Cu toate acestea, posibilitatea validității diferențiale pentru membri din diferite grupuri etnice, grupuri de gen, vorbitori nenativi de limbă engleză și alte categorii de indivizi dezavantajate în mod tradițional trebuie să fie investigată întotdeauna pentru a stabili cu certitudine faptul că scorurile utilizate în luarea deciziilor importante sunt corecte pentru toți.

O posibilă soluție a problemei capacității de predicție diferențiale a scorurilor obținute la test ar putea fi utilizarea diferitelor ecuații de regresie și a unor scoruri-prag diferite pentru selectarea indivizilor din diferite grupuri etnice și diferite categorii de gen. În cazul unor scoruri la test care supra-prezic performanța negrilor sau a hispanicilor, de exemplu, ar putea fi

nevoie de scoruri-prag mai înalte pentru membrii din aceste minorități, decât pentru albi și asiatici. Totuși, în mod evident, acest fapt este împotriva obiectivelor de extindere a oportunităților grupurilor prin acțiuni de afirmare și de creștere a diversității în învățământul superior și la locul de muncă. O altă soluție propusă, implementată în anii '80 prin General Aptitude Test Battery (GATB) construită de United States Employment Service (USES) este utilizarea normelor de sub-grup pentru a asigura rate comparabile de recomandări de angajare pentru negri, hispanici și albi. Această practică a generat o opoziție atât de mare, încât a condus la pasajul din Civil Rights Act din 1991 (P.L. 101–336) care a respins orice tip de ajustare a scorurilor bazată pe rasă, culoare, sex, religie sau origine națională. În lumina acestor obstacole, mare parte a activității din aria corectitudinii în testarea psihologică este acum concentrată pe (a) identificarea factorilor care pot contribui la predicția diferențială în grupurile etnice și de gen (vezi Steele, 1997; Willingham, Pollack, & Lewis, 2000) și (b) analizarea modului în care funcționează itemii testului pentru diferite subgrupuri cât timp testele sunt în faza de construcție, pentru a ne asigura că cei care funcționează diferit pentru diferite subgrupuri nu sunt incluși (vezi Capitolul 6).

Meta-analize

La sfârșitul anilor '70, mai multă claritate și un entuziasm reînnoit au apărut în locul perspectivei oarecum pesimiste provocată de cercetările timpurii asupra validității scorurilor testelor de selecție. Această schimbare se datorează în mare măsură utilizării meta-analizelor care permit investigatorilor să combine date provenite din numeroase studii diferite — în special din arii bogate în rezultate contradictorii — și să tragă concluzii care au

Nu uitați!

- Studiile de generalizare a validității (GV) sunt utilizate de peste un sfert de secol. Ele trebuie să fie mai numeroase și să aibă un impact mai mare asupra teoriei psihometrice și asupra practicii.
- Cititorii care doresc să urmărească subiectul GV în mai mare profunzime pot consulta *Validity Generalization:A Critical Review*, un volum editat de Murphy, Fleishman și Cleveland (2003).

autoritate mult mai mare decât cele obținute prin modalitățile tradiționale de revizuire a literaturii de specialitate. În contrast cu natura calitativă a revizuirilor literaturii realizate în mod tradițional, *meta-analizele* se bazează pe o serie de proceduri cantitative care permit sinteza și integrarea rezultatelor obținute din literatura de cercetare despre o anumită temă.

Aceste tehnici, care au fost utilizate în alte domenii științifice pentru un timp, au fost introduse în cercetarea psihometrică de Schmidt și Hunter (1977) ca o modalitate de a aborda problema generalizării validității. În ultimele două decenii, tehnicile meta-analitice au demonstrat de fapt că validitatea predictivă a scorurilor unui test nu este atât de specifică situației așa cum se credea anterior și au devenit o metodă importantă de clarificare a rezultatelor contradictorii din alte zone ale literaturii psihologice (Hunter & Schmidt, 1990, 1996).

Interesul larg răspândit pentru meta-analize și pentru utilizarea acestora a fost stimulat de realizarea faptului că multe rezultate contradictorii din cercetarea psihologică — inclusiv cele ale studiilor de validare — pot fi atribuite imperfecțiunilor studiilor individuale. Atunci când influențele unor artefacte cum ar fi eroarea de eșantionare, eroarea de măsurare, restricția asupra intervalului și dihotomizarea nejustificată a variabilelor sunt îndepărtate prin corecții statistice, apare o imagine mult mai clară.

Concomitent, se recunoaște din ce în ce mai mult faptul că testarea ipotezelor în studiile psihologice a supraevidențiat *nivelele de semnificație statistică* ce subliniază evitarea *erorilor de Tip I* (i.e. respingerea în mod incorect a ipotezei de nul ce afirmă că nu există nicio diferență atunci când această ipoteză este adevărată), neglijând posibilitatea *erorilor de Tip II* (i.e. acceptarea în mod incorect a ipotezei de nul atunci când este falsă). Din moment ce relația dintre erorile de Tip I și de Tip II este inversă, accentul pe evitarea erorilor de Tip I crește probabilitatea erorilor de tip II. În consecință, un număr enorm de rezultate ale cercetărilor care nu au atins nivelele dorite de semnificație statistică (0,05 sau 0,01), dar care, fără îndoială, pot aduce informații valoroase, au fost fie ignorate, fie excluse din literatura de specialitate. Info 5.10 enumeră câteva referințe care oferă informații suplimentare despre dezavantajele și avantajele testelor de semnificație a ipotezei de nul. Oricum, aceste discuții au condus către ceea ce majoritatea cercetătorilor consideră a fi o schimbare salutară a modului în care sunt raportate rezultatele cercetărilor. În loc să se declare pur și simplu nivelele de semnificație sau probabilitatea rezultatelor, acum se consideră necesar să se includă indicatorii *mărimii efectului* sau forța relației descoperite de un studiu de cercetare, alături de intervalele de încredere pentru mărimea efectului și pentru toate estimările parametrilor care au apărut în urma unei investigații (APA 2001).

Deși metodologia meta-analizelor este în dezvoltare, deja a adus contribuții semnificative la dovezile validității predictive a scorurilor testelor și a procedurilor care nu presupun teste — cum ar fi interviurile de angajare și inventarele de date biografice — utilizate în selectarea de personal (Hartigan & Wigdor, 1989; Schmidt & Hunter, 1998; Schmidt et al., 1993). În plus,

meta-analizele au ajutat la clarificarea literaturii de cercetare și la continua dezvoltare a teoriilor din câteva arii ale psihologiei industrial-organizaționale — cum ar fi relația dintre satisfacția în muncă și performanța în muncă — ca și în diferite alte domenii (vezi Hunter & Schmidt, 1996; Kirsch & Sapirstein, 1998; Rosenthal & DiMatteo, 2001). Un exemplu din domeniul testării pentru admiterea în învățământul superior va ilustra potențialul inerent al cercetărilor meta-analitice.

Un caz educațional ilustrativ. Studiile de validitate a scorurilor obținute la Graduate Record Examination (GRE) ca predictori ai performanței în programele universitare au o lungă istorie — datând din anii '40 — și s-au aflat continuu sub presiunea unor rezultate inconsecvente. În timp ce unele cercetări (ex.: Briel, O'Neill, & Scheuneman, 1993; Broadus & Elmore, 1983) au descoperit că testele GRE General și Subject sunt predictori ai performanțelor universitare cu o validitate destul de mare, multe altele — inclusiv câteva studii meta-analitice — au ajuns la concluzia că relația dintre scorurile obținute la GRE și diferiți indicatori ai succesului universitar era inadecvată (ex.: Goldberg & Alliger, 1992; Marston, 1971; Morrison & Morrison, 1995; Sternberg & Williams, 1997). Multe studii de validitate pentru GRE au ajuns la corelații negative ușoare până la corelații pozitive scăzute în jur de 0,20 pentru scorurile obținute la GRE la testele Verbal și Quantitative, și coeficienți ceva mai ridicați pentru scorurile la testul Subject din GRE. Deși unele dintre aceste rezultate au fost criticate pe baza artifactelor metodologice cum ar fi restricțiile asupra intervalelor atât pentru scorurile obținute la GRE, cât și pentru dimensiunile criteriilor, precum și pentru lipsa de fidelitate a criteriilor, în general, lucrările despre validitatea scorurilor obținute la GRE nu par să ofere dovezi substanțiale pentru a susține utilizarea acestor teste în luarea

Info 5.10

Surse de informații privind argumente pro și contra testelor de semnificație statistică

Timp de câteva decenii, între statisticieni și metodologi s-a defășurat o dezbatere cu privire la meritele și deficiențele inerente în utilizarea testării semnificației statistice a ipotezei de nul pentru inferențele făcute în cercetările din științele sociale. Unii dintre ei au fost atât de convinși că aceste teste au un efect dăunător asupra întreprinderii științifice, încât au propus interzicerea apariției testelor de semnificație în rapoartele de cercetare. Deși nu a fost instituită o astfel de interdicție, este acum obișnuit ca revistele de psihologie și revistele din multe alte discipline relaționate, să ceară estimări ale mărimii efectului ori de câte ori sunt raportate valori ale probabilității (p), împreună cu intervale de încredere pentru mărimile efectului, coeficienți de corelație și alte estimări ale parametrilor populației. Informații suplimentare despre aceste aspecte implicate în dezbatere pot fi găsite în următoarele surse:

- Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15.
- Cohen, J. (1994). The earth is round ($p < 0,05$). *American Psychologist, 49,* 997–1003.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 25–32.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

deciziilor de admitere în învățământul universitar (vezi Kuncel, Campbell, & Ones, 1998).

Pe acest fond, Kuncel, Hezlett, și Ones (2001) au condus recent o meta-analiză meticuloasă și completă a datelor GRE din 1 753

de eșantioane independente, cuprinzând în total 82 659 de absolvenți de facultate. Acest studiu s-a adresat în mod sistematic aspectelor teoretice, statistice și metodologice ale literaturii despre validitatea predictivă a scorurilor obținute la GRE. Kuncel și colegii săi au examinat relația dintre cinci predictori — scorurile la testul GRE Verbal (GRE-V), Cantitativ (GRE-Q), Analitic (GRE-A) și Subiect, ca și media generală obținută în facultate (UGPA) — și opt criterii diferite ale succesul în învățământul universitar, inclusiv GPA (punctajul mediu anual) în primul an și GPA la absolvire (GGPA) (media generală), scoruri ale examinării comprehensive, evaluări ale eșecurilor, nivelul achizițiilor de cunoștințe și indicatori numerici legați de productivitatea cercetării. Ei au condus analize separate pentru eșantionul total și pentru sub-eșantioane care reprezintă studenți din domeniile umaniste, din științe sociale, științele biologice și fizico-matematice, precum și pentru vorbitorii nenativi de limbă engleză și pentru studenții cu vârstă mai mare decât vârsta tradițională a absolvirii. Printre multe rafinamente metodologice, Kuncel și colegii săi au utilizat în analizele lor corecții pentru restricțiile asupra intervalului și pentru schimbările în variabilitatea distribuțiilor variabilelor predictor și criteriu, ca și pentru lipsa de fidelitate a dimensiunilor criteriului. În plus, acești cercetători au adus în discuție un număr de capcane potențiale, inerente în meta-analize.

Kuncel și colegii săi (2001) au raportat rezultatele majore obținute în termenii corelațiilor medii observate, alături de abaterile standard, ponderate pentru dimensiunea eșantionului. De asemenea, au raportat și validitățile operaționale estimate cu abaterile lor standard și intervale de încredere de 90%. Aceste rezultate indică faptul că cele patru instrumente GRE (testele GRE-V, GRE-Q, GRE-A și GRE Subiect) sunt predictori suficient de buni pentru cele mai multe criterii utilizate pentru eșantionul total, precum și pentru sub-eșantioane. De fapt, în cele mai multe cazuri, scorurile GRE par a fi predictori mai buni decât UGPA. Scorurile obținute la testul GRE Subiect s-au dovedit a fi cei mai buni predictori ai punctajului mediu al absolvenților pentru toate disciplinele, cu validități operaționale estimate cuprinse între 0,40 și 0,49. Prin contrast, scorurile la testele generale (GRE-V, GRE-Q și GRE-A) — deși corelează substanțial cu cele de la testele Subiect — au avut coeficienți de validitate operațională cuprinși între 0,27 și 0,48. Kuncel și colaboratorii săi au ajuns la concluzia că, deși scorurile obținute la testul general GRE contribuie doar într-o mică măsură la validitate atunci când sunt adăugate scorurilor testului Subiect, ele pot fi totuși valoroase, mai ales pentru studenții care au absolvit o facultate dintr-un alt domeniu decât cel pentru care candidează în momentul testării.

Pe scurt, meta-analiza realizată de Kuncel și colegii săi (2001) sugerează în mod clar (a) că mare parte din inconsistența studiilor anterioare de validare a scorurilor obținute la GRE a fost rezultatul restricției impuse asupra intervalului și erorii de eșantionare din acele studii și (b) că scorurile obținute la GRE merită să aibă un rol în procesul de admitere în învățământul superior. Totuși, acești autori nu investighează aspectele predicției diferențiale pentru femei și membri ai minorităților etnice sau rasiale și admit faptul că există încă posibilitatea de îmbunătățire a validității procesului de admitere în învățământul superior. Cu privire la ultimul punct, următoarele noțiuni trebuie reținute:

- Scopul scorurilor GRE și a majorității celorlalți predictori utilizați în deciziile de selecție nu este de a estima poziția

exactă a candidaților din perspectiva criteriului, ci de a determina dacă aceștia pot atinge nivelul necesar de succes. Dacă este necesară o predicție exactă a scorurilor la criteriu, trebuie într-adevăr să se ia în considerare limitele erorii (SE_{est}) pentru coeficienți cuprinși între 0,30 și 0,40.

- Performanța la cele mai multe criterii, inclusiv succesul absolvirii, este determinată de factori multipli, printre care se numără caracteristicile atitudinale și emoționale și obiceiurile comportamentale, precum și talentele creative și practice care nu sunt măsurate prin GRE sau alte teste cognitive utilizate de obicei ca predictori.
- Multe alte decizii de selecție, inclusiv admiterile în învățământul superior, sunt rareori luate numai pe baza unui singur predictor. Prin urmare, aspectul deosebit de important al testelor de selecție privește utilitatea lor. Aceasta înseamnă că întrebarea care trebuie pusă este dacă utilizarea scorurilor la test ca parte a procesului de luare a deciziilor crește numărul de decizii valide peste ceea ce s-ar obține prin utilizarea predictorilor non-test, cum ar fi GPA. În cele mai multe situații, inclusiv în cazul deciziilor de admitere din învățământul superior, datele sugerează că scorurile la test contribuie la eficiența predictivă în luarea deciziilor (vezi Hartigan & Wigdor, 1989; Kobrin, Camara, & Milewski, 2002; Kuncel, Hezlett, & Ones, 2001).

Nu uitați!

Numeroase informații cu privire la validitatea predictivă a testelor utilizate în învățământul superior, inclusiv multe dintre aspectele legate de predicția diferențială pentru diferite subgrupuri, pot fi găsite pe internet. Vezi, mai ales, următoarele site-uri:

- ACT (http://www.act.org)
- The College Board (http://www.college.board.com)
- Educational Testing Service (ETS; http://www.ets.org)

Dincolo de selecție: utilizarea scorurilor la test pentru alte tipuri de decizii

Până acum, prezentarea procedeelor validării în raport cu criteriul s-au centrat în primul rând asupra utilizării testelor pentru selecția realizată fie pe baza validității predictive a scorurilor obținute la test, fie pe baza validității concurente. Deciziile de *selecție* sunt cele care necesită o alegere între două alternative. În cadrul ocupațional și în cel educațional, alternativele obișnuite sunt fie acceptarea, fie respingerea unui candidat; în cadrul clinic și în cel judiciar, deciziile de selecție presupun de obicei determinarea prezenței sau absenței unui sindrom sau a unei condiții particulare. Termenul *triere* se referă la un pas preliminar al unui proces de selecție care se desfășoară de obicei pentru a separa indivizii care merită sau necesită o evaluare mai extinsă de ceilalți care nu se află în această situație. De exemplu, periodic, multe clinici utilizează un scurt chestionar pentru a tria persoanele cu tulburări cum ar fi depresia sau anxietatea din populația generală; angajatorii pot tria candidații cu ajutorul unui instrument scurt pentru a limita lotul de candidați la cei care îndeplinesc cerințele minime pentru postul respectiv. Scorurile obținute la testele psihologice sunt utilizate și pentru decizii de plasament și clasificare, ambele presupunând mai mult de două opțiuni. Dintre acestea, deciziile de *repartizare* sunt cele mai simple. Ele presupun distribuirea indivizilor în categorii sau tratamente separate pe baza unui singur scor sau a unui scor combinat calculat cu ajutorul unei singure ecuații de regresie, în raport cu un singur criteriu. Deși deciziile de repartizare nu presupun opțiunea respingerii indivizilor care nu ating un anumit nivel de performanță la un test sau la un predictor, ele nu sunt substanțial

diferite de deciziile de selecţie în termenii dovezilor pe care le cer, o relaţie demonstrabilă între unul sau mai mulţi predictori şi un criteriu. De exemplu, scorurile la un test de citire pot fi utilizate pentru a repartiza elevii în grupe instrucţionale adecvate nivelelor lor de abilităţi de citire. În mod similar, scorurile la o scală de depresie pot fi utilizate pentru a clasifica pacienţii de psihiatrie în termenii gravităţii simptomelor lor depresive în scopul determinării tipurilor şi nivelurilor adecvate de intervenţie terapeutică.

Pe de altă parte, deciziile de *clasificare* sunt cu mult mai complicate. În clasificare — ca şi în repartizare — nimeni nu este respins, dar indivizii trebuie să fie distribuiţi *diferenţiat* în categorii sau tratamente distincte pe baza unui criteriu multiplu. Aceasta înseamnă că este nevoie de predictori multipli şi că relaţia acestora cu fiecare criteriu trebuie să fie determinată în mod independent, prin ecuaţii de regresie separate. Cel mai potrivit instrument pentru deciziile de clasificare este o baterie de teste sau predictori ale căror rezultate sunt validate în raport cu diferite criterii pentru care se fac predicţii şi apoi sunt combinate în ecuaţii care reflectă ponderea lor relativă în predicţia fiecărui criteriu.

Deciziile de clasificare sunt cerute în cadrul ocupaţional, educaţional, clinic şi în consilierea psihologică. În domeniul ocupaţional, inclusiv în cadre militare sau industriale, aceste decizii sunt necesare atunci când sunt evaluate aptitudinile unui lot de personal disponibil pentru a distribui indivizii la locurile de muncă sau în programele de instruire în care este cel mai probabil să funcţioneze eficient. De asemenea, consilierea vocaţională a indivizilor care doresc să se decidă în privinţa unui program de studiu sau în alegerea unei cariere necesită decizii de clasificare. În cadrul clinic deciziile de clasificare trebuie luate în cazuri în care este

nevoie de diagnostic diferenţial. Un exemplu tipic ar fi nevoia de a stabili dacă un pacient mai în vârstă care prezintă simptome de depresie şi probleme de memorie suferă datorită unei tulburări depresive care influenţează memoria şi concentrarea, datorită unui proces incipient de demenţă care cauzează depresie sau e prezentă o combinaţie a celor două condiţii.

Bateriile de teste utilizate pentru deciziile de clasificare trebuie evaluate pe baza dovezilor validităţii diferenţiale. În acest context, termenul *validitate diferenţială* înseamnă că o baterie trebuie să fie capabilă să prezică sau să stabilească diferenţele pentru două sau mai multe criterii. În cazul unei probleme de clasificare cu criteriu dublu, o baterie ideală ar consta din predictori care corelează puternic cu un criteriu şi nu corelează sau corelează negativ cu celălalt criteriu. De exemplu, în problema diagnosticului diferenţial depresie versus demenţă am putea căuta diferenţe în secvenţele temporale ale simptomelor de depresie şi deteriorare cognitivă sau diferenţe în nivelele relative de performanţă la diferite tipuri de teste de memorie.

Atunci când clasificarea presupune predicţii în raport cu mai mult de două criterii, cum ar fi distribuirea personalului în unul dintre câteva locuri de muncă posibile sau programe de instruire, problema stabilirii dovezilor validităţii devine chiar mai complexă. În acest tip de situaţie, un predictor care corelează la fel de mult cu toate criteriile implicate în decizie — cum ar fi un test de inteligenţă generală ce are legătură cu cele mai multe criterii de performanţă în muncă — are o utilitate relativ redusă. O modalitate posibilă de a aborda problemele de clasificare de acest tip este utilizarea analizelor funcţionale discriminatorii multiple. *Funcţiile discriminatorii* presupun aplicarea combinaţiilor ponderate ale scorurilor la predictori — derivate prin intermediul

analizelor regresiei — pentru a determina cât de mult se potrivește profilul scorurilor unui individ cu profilul indivizilor din diferite grupuri ocupaționale, din diferite specialități sau din diferite categorii de diagnostic psihiatric. Deși funcțiile discriminatorii sunt utile în anumite situații (atunci când criteriile constau pur și simplu din apartenența la un anumit grup versus altul sau când există o relație neliniară între un criteriu și unul sau mai mulți predictori), acestea sunt neadecvate cerințelor din numeroase situații pentru că nu permit predicția nivelului reușitei într-un domeniu specific. Pentru un exemplu de aplicare a analizei funcționale discriminatorii în diferențierea profilelor WAIS-R pentru pacienți cu traumatism cranian incontestabil de cele ale subiecților instruiți să pretindă că au un traumatism cranian vezi Mittenberg, Theroux-Fichera, Zielinski, și Heilbronner (1995).

O altă stategie tradițională care poate fi utilizată atât pentru problemele de selecție, cât și pentru cele de clasificare este *validarea sintetică* (Balma, 1959). În mod esențial, această tehnică se bazează pe analize detaliate ale muncii care identifică elementele componente specifice ale muncii și ponderea lor relativă în diferite ocupații. Pe baza unor astfel de analize, coeficienții de regresie pentru scorurile testului stabiliți anterior care fac predicții în legătură cu acele elemente separate ale muncii pot fi combinați într-o nouă baterie, sintetică, ce va duce la predicții ale performanței la sarcinile de muncă aflate în chestiune. Procedurile statistice asociate cu această metodă au fost introduse de Primoff (1959; Primoff & Eyde, 1988) și au fost de atunci extinse de alții. Cu toate acestea, pentru a fi utile în deciziile de clasificare, strategiile de validare sintetică trebuie să implice predictori care prezintă o bună validitate discriminatorie, în afară de cazul în care

chiar componentele criteriului corelează substanțial (Guion, 1998, pp. 354–355).

Perspective suplimentare asupra validității în raport cu criteriul

Câteva progrese metodologice prezentate anterior în acest capitol, cum ar fi modelarea ecuației structurale și studiile meta-analitice de generalizare a validității au adus în cercetările recente problemele validării în raport cu criteriul. În același timp, disponibilitatea acestor instrumente statistice din ce în ce mai sofisticate a atras atenția atât asupra predictorilor, performanței la criteriu, precum și asupra interrelației dintre acestea (vezi J. P. Campbell, 1990). Unul dintre cele mai semnificative progrese în termenii predictorilor este recunoașterea faptului că includerea factorilor legați de dimensiunile personalității — alături de cei legați de aptitudini — poate crește validitatea instrumentelor utilizate pentru predicția performanțelor în numeroase domenii (vezi Lubinski & Dawis, 1992). Cu privire la problema criteriului, natura multifațetată a performanței în cele mai multe sarcini de muncă și medii educaționale este acum larg recunoscută și cercetată din ce în ce mai mult. Aceasta necesită analizarea diferitelor elemente care conduc la succes, stabilirea valorii lor relative, recunoașterea celor care se află sub controlul individului, dezvoltarea unor metode pentru a evalua separat fiecare element și apoi combinarea lor într-un instrument de măsurare totală a performanței care ține cont de toți acești factori și de mărimea lor relativă.

Un model de program pentru validarea legată de criteriu

Un exemplu bun al aplicării numeroaselor inovații din metodele de cercetare de validare poate fi găsit în activitatea pe care John P. Campbell și colaboratorii săi (J. P. Campbell, 1990, 1994; J. P. Campbell & Knapp, 2001) au desfășurat-o în ultimele două decenii, alături de efortul de a evalua și îmbunătăți procedurile utilizate în U.S. Army pentru selecția și clasificarea personalului, cunoscută drept Proiectul A. O relatare completă a acestei activități, care e posibil să fie cel mai mare proiect din istoria cercetărilor din domeniul ocupațional, nu este posibilă în acest context. Totuși, câteva dintre punctele sale importante sunt prezentate aici pentru a le oferi cititorilor o idee despre anvergura și semnificația lui.

Proiectul A

Cercetătorii din cadrul Proiectului A au utilizat o bază de date colectate de la mai mult de 50 000 de oameni și au selectat 21 de specialități ocupaționale militare (SOM) dintr-un total de peste 200 de posturi din cadrul armatei, pe care au realizat studii concurente și longitudinale de validare predictivă. Predictorii studiați au inclus cele 10 subteste din Armed Services Vocational Aptitude Battery (ASVAB) enumerate în Info 5.11. Diferite scoruri compuse constând din diferite combinații ale subtestelor din ASVAB au fost utilizate în mod tradițional pentru selecția și clasificarea indivizilor care candidau pentru a servi în Armed Forces. În plus, câteva noi instrumente — printre care teste de aptitudini psihomotorii și spațiale ca și instrumente de măsurare a personalității și intereselor — au fost incluse în activitatea de validare.

Info 5.11

Armed Services Vocational Aptitude Battery (ASVAB)

ASVAB este administrată ca test computerizat adaptativ (CAT-ASVAB) în unitățile militare de procesare a admiterilor și ca test creion-hârtie în zonele mobile de testare. Indivizii cărora le este administrat ASVAB au fost deja pretestați cu instrumente de triere mai scurte, administrate de recrutori. Cele 10 subteste care alcătuiesc bateria ASVAB, utilizată de toate unitățile U.S. Armed Forces pentru selecția și clasificarea personalului sunt următoarele:

- Raționament aritmetic
- Operații numerice
- Înțelegerea unui paragraf
- Cunoașterea semnificației cuvintelor
- Viteză de codare
- Cunoștințe științifice generale
- Cunoștințe matematice
- Informație electronică
- Înțelegere mecanică
- Informații despre autovehicule

Primele patru subteste alcătuiesc compozitul Armed Forces Qualification Test (AFQT) care, împreună cu un statut înalt al absolvirii liceului, este utilizat pentru a determina eligibilitatea pentru înrolare. Subtestele ASVAB sunt combinate și în 10 scoruri compuse pentru aptitudini diferite care sunt utilizate ca predictori ai succesului în programele de instruire din școala militară (J. P. Campbell & Knapp, 2001, pp. 14–18).

Cercetătorii din Proiectul A au realizat analize extensive ale sarcinilor de muncă și au acordat o atenție specială standardizării instrumentelor de măsurare a productivității muncii pentru fiecare poziție SOM. Au fost utilizate analize factoriale exploratorii

sau confirmatorii, ca și alte tehnici statistice în studiile de modelare a performanței care au condus la specificarea scorurilor factorilor utilizate drept criterii pentru trei stadii diferite ale carierei: performanța la sfârșitul perioadei de instruire, performanța la locul de muncă în primul stagiu și în al doilea stagiu. Explorarea dovezilor validității în diferite stadii ale carierei a fost un aspect semnificativ al acestei cercetări deoarece, pe lângă construirea unei baterii de teste capabilă de predicție diferențială, un alt obiectiv important al sarcinilor de gestionare a personalului din All Volunteer Armed Forces este acela de a minimiza retragerile.

Analizele de validare concrete realizate pe durata Proiectului A au inclus (a) corelații între fiecare predictor și performanța la criteriu; (b) comparații între validitatea incrementală graduală oferită de măsurătorile experimentale cu privire la fiecare criteriu, peste și sub puterea predictivă a scorurilor ASVAB disponibile anterior; (c) dezvoltarea și compararea diferitelor ecuații optime pentru validitatea maximă în predicția longitudinală a performanței din primul stagiu și (d) analize ale validității ecuațiilor alternative utilizând diferite combinații de date ale testului și performanțele anterioare pentru predicția performanței din al doilea stagiu. Generalizarea validității și metode de validare sintetică au fost utilizate pentru a investiga eficiența diferitelor combinații de predictori pentru cele 21 de poziții SOM selectate inițial, pentru analize și pentru multe alte SOM sau clustere de ocupații. Studii suplimentare au inclus o evaluare a utilității și beneficiilor obținute prin strategiile și condițiile de clasificare alternativă. Printre numeroasele descoperiri utile din Proiectul A, una dintre cele mai semnificative a fost coroborarea valorii potențiale a câtorva noi instrumente studiate, inclusiv a celor legate de dimensiuni ale personalității, ca și a câtorva teste noi ale aptitudinilor

Nu uitați!

Judecățile cu privire la validitatea scorurilor unui test sunt relative. Atunci când dovezile cu privire la validitatea scorurilor produse de un test care s-au acumulat sunt luate în considerare în termeni abstracți, ele pot sau nu pot fi considerate ca suficiente pentru scopul urmărit de test.

Cu toate acestea, orice scor sau orice set specific de scoruri provenite de la un individ sau de la un grup este luat în considerare, trebuie să se înțeleagă faptul că scorurile pot fi influențate de factori unici, relevanți doar pentru persoanele testate, pentru examinator, pentru contextul în care are loc testarea și de interacțiunea dintre acești factori. Astfel, când sunt interpretate scorurile unui test, trebuie să se țină cont întotdeauna de situația de testare.

În plus, inferențele făcute pe baza scorurilor obținute la un test necesită informații cu privire la cadrul de referință și fidelitatea lor și dovezi ale validității din toate sursele relevante.

psihomotorii și spațiale. Acești predictori suplimentari se află acum în diferite stadii preliminare de implementare și sunt explorați în continuare. Conform principalilor cercetători din cadrul Proiectului A, aceste instrumente nu sunt încă deplin operaționale din cauza obstacolelor puse de numeroase părți implicate în sistemul de selecție și clasificare din armată, precum și a celor generate de inerția organizațională (J. P. Campbell & Knapp, 2001, pp. 570–574). Cu toate acestea, Proiectul A și investigațiile ulterioare pe care le-a generat au adus deja numeroase contribuții substanțiale și progrese metodologice care, fără îndoială, îmbunătățesc calibrul cercetărilor de validare pentru selecție și clasificare atât în cadrul militar, cât și în numeroase alte contexte și astfel ne vor ajuta în urmărirea obiectivului de maximizare a utilizării talentelor umane.

Aspecte suplimentare ale validității: utilitate și consecințe

Există două aspecte semnificative care privesc utilizarea scorurilor obținute la un test și care sunt strâns legate de validitatea lor, dar nu sunt legate în mod necesar de esența acesteia, și anume, utilitatea și consecințele utilizării acestor teste. Complexitatea acestor teme ne împiedică să facem o prezentare extinsă a lor în cadrul acestui volum. Totuși, importanța lor crucială pentru testarea psihologică justifică introducerea lor în acest moment, urmată de o prezentare mai extinsă în Capitolul 7.

Evaluarea utilității testării psihologice

Utilitatea testelor și a scorurilor obținute la teste se referă la beneficiile pe care acestea le aduc în luarea deciziilor. Utilitatea este contingentă măsurii în care utilizarea testelor poate spori rata de precizie a inferențelor și deciziilor pe care dorim să le facem — suplimentară celei pe care am obține-o dacă am utiliza alte instrumente disponibile. În mod tipic, utilitatea este evaluată în termeni economici, cum ar fi rata costuri-beneficii implicată în utilizarea datelor care provin din teste versus date care nu provin din teste. Dat fiind faptul că întotdeauna utilizarea unui test are loc în cadrul unui context, analiza costurilor și a beneficiilor sale trebuie să ia în calcul, în mod necesar, datele suplimentare relevante pentru fiecare situație particulară în care se urmărește utilizarea testelor. În funcție de context, aceste date includ aspecte ca probabilitățile și riscurile presupuse de realizarea unor determinări fals pozitive sau fals negative, existența unor instrumente alternative — și eficiența lor relativă în comparație cu testele — ca și relativa ușurință sau dificultate a determinărilor care trebuie făcute. Acest aspect al utilizării testelor este o parte a unei teme mai mari a teoriei deciziei, relevantă nu numai pentru psihometrie în particular și psihologie în general, dar și pentru diferite alte domenii ca medicina, economia, jurisprudența, științe militare și jocuri, precum și pentru orice alte medii umane în care este cerută planificarea strategică. Pentru exemple ale numeroaselor contribuții pe care experții în psihometrie le-au adus teoriei deciziilor în ultimele câteva decenii vezi Boudreau (1991), Brogden (1946), Brown și Ghiselli (1953), Buchwald (1965), Cronbach și Gleser (1965), Hunter și Schmidt (1981), Schmidt, Hunter, McKenzie, & Muldrow (1979) și Taylor și Russell (1939). O prezentare excelentă a aspectelor statistice ale teoriei deciziei în contextul sporirii preciziei și utilității deciziilor de diagnostic este disponibilă într-un raport recent realizat de Swets, Dawes și Monahan (2000).

Evaluarea consecințelor utilizării testelor

Consecințele individuale și sociale ale utilizării testelor, care pot fi pozitive sau negative, trebuie evaluate în termenii valorii implicațiilor lor. Anumiți teoreticieni, mai ales Messick (1989,1995), au susținut faptul că judecățile cu privire la validitate sunt de fapt judecăți de valoare. Messick a propus includerea aspectelor *consecințelor* utilizării și interpretării scorurilor unui test — adică evaluarea consecințelor lor sociale intenționate și neintenționate — în cadrul noțiunii de validitate în sens larg. Criticând propunerea lui Messick, unii autori au etichetat-o drept „validitate a consecințelor", un termen pe care Messick nu l-a utilizat niciodată. Oricum, acest aspect particular al contribuțiilor imense ale lui Messick la teoria și practica psihometriei nu a fost larg adoptat

în testarea psihologică. Cei mai mulți susțin că, deși aspectele legate de consecințe sunt de o deosebită importanță și ar trebui investigate înaintea implementării și apoi ar trebui documentate, ele se situează în domeniul eticii profesionale, al valorilor morale și considerațiilor politice și nu țin de determinarea validității în sine (Cole & Moss, 1989; Lees-Haley, 1996; Linn, 1998).

De exemplu, *Ethical Principles of Psychologists and Code of Conduct* (APA, 2002), cere tuturor psihologilor — inclusiv celor care utilizează teste și alte instrumente de testare — să țină cont de posibilele ramificații ale activității lor pentru a maximiza beneficiile și a minimiza daunele aduse altora. Mai ales utilizatorii testelor și altor instrumente de evaluare sunt obligați să nu interpreteze greșit scorurile, să nu le utilizeze greșit și să obțină consimțământul subiecților evaluați înaintea testării cu privire la scopurile testării, maniera în care vor fi utilizate scorurile, persoanele care vor avea acces la scoruri și alte aspecte de acest tip. Constrângeri similare se aplică activității autorilor de teste, celor care le publică sau le revizuiesc, ca și altor profesioniști implicați în testarea psihologică (Society for Industrial and Organizational Psychology [SIOP], 2003). Dacă s-ar adăuga noțiunii deja complexe de validare a scorurilor unui test fațeta suplimentară a evaluării ramificațiilor utilizării testelor în termenii unor preocupări sociale mai largi — cum ar fi echilibrarea principiilor morale ca dreptate și corectitudine — s-ar pune pe umerii profesioniștilor din testarea psihologică o povară exagerată, care aparține societății ca întreg.

Un exemplu al modului în care principiile și praticile etice din testare se pot aplica în domeniul educației poate fi găsit într-un ghid de resurse pentru profesori care prezintă atât standardele profesionale, cât și principiile relevante pentru utilizarea testelor

în scopul luării unor decizii educaționale deosebit de importante pentru studenți (U.S. Department of Education, Office for Civil Rights, 2000). Seturi similare de recomandări pentru utilizarea testelor în alte domenii specializate vor fi discutate în Capitolul 7 (SIOP, 2003).

Comentarii de încheiere

Info 5.12 subliniază modul în care diferite surse de dovezi și strategii de validare susțin interpretarea scorurilor provenite de la un singur test, în funcție de scopurile pentru care este utilizat testul. În cele mai multe cazuri, în măsura în care interpretarea propusă pentru scorurile unui test se îndepărtează de scopul inițial pentru care a fost construit testul, dovezile pentru utilizări și interpretări alternative vor deveni mai puțin directe. De exemplu, testul utilizat în Info 5.12 a fost examenul final la cursul de Calcul I; asfel, putem presupune că scopul său inițial a fost de a determina dacă subiecții testați stăpânesc suficient de bine conținutul acelui curs pentru a obține o notă de trecere. În măsura în care interpretarea scorurilor la acest test este extinsă la determinarea pregătirii necesare pentru Calcul II, pentru predicția succesului în specializarea „matematică" sau la utilizarea scorurilor obținute la test ca înlocuitor al aptitudinii matematice în investigarea corelatelor tipurilor de personalitate, legătura dintre dovezile validității și interpretarea intenționată devine din ce în ce mai slabă. Aceasta nu înseamnă că examenul final la Calcul I nu ar trebui utilizat în alte scopuri decât cel inițial, ci sugerează că este nevoie de dovezi suplimentare pentru aceste scopuri suplimentare.

În acest capitol am discutat dovezile validității scorurilor obținute la test din punctul de vedere al testului ca întreg. În

Info 5.12

Strategiile de validare legate de interpretarea scorurilor unui test

Testul ale cărui scoruri vor fi interpretate	Scopul propus al interpretării scorurilor testului	Tipul de strategie de validare dorită	Posibile surse de dovezi
Examenul final la Calcul I	Să se determine dacă studenții pot obține notă de trecere la cursul Calcul I	Validitate de conținut	Relevanța și reprezentativitatea conținutului testului în raport cu materia acoperită de cursul Calcul I
	Să se determine dacă studenții sunt pregătiți pentru Calcul II	Validitate în raport cu criteriul, de tip concurent	Corelație pozitivă înaltă între scorurile la testul Calcul I și notele obținute la cursul Calcul II
	Predicția succesului studenților ca matematicieni	Validitate în raport cu criteriul, de tip predictiv	Corelație pozitivă înaltă între scorurile obținute la testul Calcul I și încheierea modulului de matematică
	Investigarea relației dintre aptitudinile matematice și tipul de personalitate	Convergență	Susținerea ipotezei că studenții introverți vor obține scoruri mai înalte decât studenții extraverți la testul Calcul I.

următorul capitol ne vom îndrepta atenția către o analiză mai minuțioasă a datelor testului din perspectiva unităților de eșantioane de comportament care alcătuiesc scorurile testului, și anume itemii unui test.

Testează-te singur

1. ***Validitatea* este măsura în care**
   a. un test măsoară ceea ce își propune să măsoare.
   b. dovezile susțin inferențele făcute pornind de la scorurile la test.
   c. scorurile la test sunt constante în diferite situații.

2. **În ultimele decenii, diferite forme ale dovezilor validității au fost subsumate noțiunii de *validitate* _____.**
   a. *de conținut*
   b. *concurentă*
   c. *predictivă*
   d. *de construct*

3. **Intervalul nomotetic se referă la**
   a. o rețea de relații între instrumente.
   b. descompunerea sarcinilor.
   c. identificarea diferențelor dintre subiecții testați.
   d. aria constructului măsurat.

4. **Dovezile validității bazate pe conținutul testului și procesele de răspuns se aplică în special la**
   a. inventarele de interese.
   b. testele educaționale.
   c. teste de personalitate.

5. ***Validitatea de aspect* se referă în principal la**
   a. reprezentativitatea conținutului testului.
   b. dovezi ale validității din perspectivă psihometrică.
   c. caracteristici superficiale ale unui test.
   d. volumul de date empirice de validare acumulate pentru un test.

6. **Pentru a colecta dovezi ale validității discriminatorii, se vor corela scorurile obținute la teste care își propun să evalueze constructe _____.**
   a. identice
   b. similare, dar nu identice
   c. diferite

7. **Unul dintre cele mai utile aspecte ale analizei factoriale aplicate în cercetarea validității unui test este faptul că rezultatele aplicării acestei tehnici pot**
   a. simplifica interpretarea și raportarea scorurilor unui test.
   b. revela aspectele esențiale ale constructelor pe care testele le evaluează.
   c. pot fi generalizate la alte populații.

8. **Care dintre următoarele afirmații despre dimensiunile criteriului *nu* este adevărată?**
   a. Dimensiunile criteriului pot diferi în termenii fidelității și validității lor.
   b. Diferite dimensiuni ale criteriului nu corelează întotdeauna una cu celălaltă.
   c. Dimensiunile criteriului pot sau nu să fie generalizate la diferite grupuri.
   d. Cele mai bune dimensiuni ale criteriului sunt disponibile de obicei în momentul testării.

9. **Erorile standard ale estimării sunt utilizate pentru a stabili**
   a. fidelitatea criteriului.
   b. fidelitatea predictorilor.
   c. precizia scorurilor obținute.
   d. precizia cu care sunt prezise criteriile.

10. **Din punctul de vedere al procedurilor de validare în raport cu criteriul, care dintre următoarele tipuri de decizii este cel mai complex?**
    a. Selecția
    b. Repartizarea
    c. Clasificarea.

Răspunsuri: 1. b; 2. d; 3. a; 4. b; 5. c; 6. c; 7. a; 8. d; 9. d; 10. c

Considerații esențiale despre itemi

Itemii unui test sunt unitățile care alcătuiesc un test și mijloacele prin care sunt acumulate eșantioane ale comportamentului subiectului testat. Prin urmare, calitatea globală a unui test depinde în primul rând de calitatea itemilor care îl compun, deși numărul de itemi dintr-un test și secvențialitatea sau poziția lor în cadrul testului sunt de asemenea aspecte de importanță fundamentală. Așa cum testele sunt evaluate cu privire la măsura în care îndeplinesc scopurile propuse, itemii individuali trebuie să fie evaluați pe baza măsurii în care urmăresc scopurile pe care le are testul ca întreg. *Analiza de item* este un termen general care se referă la toate tehnicile utilizate pentru a evalua caracteristicile itemilor unui test și calitatea acestora în timpul procesului de dezvoltare și construire a testului.

Analiza de item presupune proceduri calitative și cantitative. Precedurile *analizei de item calitative* se bazează pe judecata evaluatorilor cu privire la caracteristicile concrete și stilistice ale itemilor, precum și pe precizia și corectitudinea acestora. Criteriile majore utilizate pentru a evalua calitativ itemii sunt (a) adecvarea conținutului și formatului itemului la scopul testării și la

populația căreia îi este destinat testul, (b) claritatea exprimării, (c) corectitudinea gramaticală și (d) respectarea anumitor reguli de bază pentru scrierea itemilor care au apărut de-a lungul timpului. Așa cum vom discuta mai târziu în acest capitol, de asemenea, conținutul itemului este examinat cu atenție pentru a identifica și elimina posibilele surse de prejudecăți sau portretizări ofensatoare ale unui subgrup specific al populației. Info 6.1 enumeră lucrări care oferă informații despre procesul de elaborare a itemilor și despre regulile practice de scriere a itemilor unui test. *Analiza de item cantitativă* presupune diferite procedee statistice destinate să stabilească caracteristicile psihometrice ale itemilor pe baza răspunsurilor obținute de la eșantioanele utilizate în procesul de construire a testului. Cea mai mare parte din acest capitol se ocupă de analiza cantitativă a itemilor unui test.

Contextul analizei de item: construirea testului

În mod fundamental, procedurile implicate în generarea itemilor, selecția itemilor și analiza de item sunt legate de tema teoriei și proiectării unui test. Prin urmare, ele au o importanță deosebită pentru autorii testelor. Utilizatorii testelor trebuie să fie familiarizați cu aceste proceduri pentru a înțelege natura sarcinii presupuse de un test și a evalua instrumentele pe care le selectează. Cu toate acestea, în momentul în care un test este disponibil utilizatorilor, el este deja un produs finit. Din punctul de vedere al utilizatorilor testului, din momentul în care au selectat un test, itemii acestuia prezintă interes — mai ales în contextul evaluării individuale — în primul rând ca mijloace de a observa și investiga răspunsurile subiectului testat din perspectiva unică a situației și circumstanțelor specifice în care este administrat un test. În

Info 6.1.

Scrierea itemilor unui test

Pentru a înțelege procesul de pregătire a itemilor pentru testele de aptitudini, precum și pentru o ghidare explicită a modului de scriere a acestora, cititorii pot consulta una dintre următoarele surse:

- Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Erlbaum.
- Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn & Bacon.
- Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (ediția a 2-a). Mahwah, NJ: Erlbaum.

Deși nu există ghiduri comparabile pentru aria vastă a abordărilor construirii instrumentelor de evaluare a personalității, unele principii de bază pentru pregătirea itemilor obiectivi pot fi găsite în următoarele cărți:

- Aiken, L. R. (1996). *Rating scales and checklists: Evaluating behavior, personality and attitudes*. New York: Wiley.
- Aiken, L. R. (1997). *Questionnaires and inventories: Surveying opinions and assessing personality*. New York: Wiley.
- Fink , A. (2002). *How to ask survey questions* (ed. a 2-a, Vol. 2).Thousand Oaks, CA: Sage. [este unul dintre cele zece volume din *Survey Kit*, editat de Arlene Fink și publicat de Sage.]

evaluarea individuală, modurile în care examinații răspund la sarcinile testului și tiparele lor particulare de răspuns pot oferi informații suplimentare, care se adaugă procesului de interpretare a scorurilor obținute la test. Bineînțeles, utilizatorii sunt preocupați, de asemenea, de aspectele practice ale itemilor testului.

Printre aceste aspecte, cel mai important este adecvarea itemilor la tipuri specifice de cadre și examinați, ușurința cu care itemii pot fi administrați și scorați, timpul presupus de administrare a itemilor și gradul de instruire necesar pentru a stăpâni procedurile implicate în administrarea și scorarea itemilor.

Procedeele analizei de item sunt implementate în diferite momente ale procesului de construire a testului, un proces care include alți câțiva pași. Pentru a oferi un context prezentării analizei de item, în următoarele câteva paragrafe sunt descriși pe scurt pașii presupuși de construirea unui test. Prezentări mai extinse ale procesului de construire a unui test sunt disponibile în numeroase surse (e.g., AERA, APA, & NCME, 1999, cap. 3; DeVellis, 2003; Ramsay & Reynolds, 2000; Robertson, 1992). Așa cum specifică Robertson (1992), elaborarea unui test standardizat presupune o investiție considerabilă de timp și bani, care cere o expertiză profesională specializată în psihometrie și în domeniile specifice de care se ocupă testul. Din acest motiv, construirea testelor destinate comercializării — spre deosebire de instrumentele experimentale ce urmează să fie utilizate în principal în scopul cercetării, testele utilizate la clasă și cele construite de angajatori pentru a fi folosite local — este desfășurată de obicei de firme care publică teste, care au resursele financiare necesare și dispun de expertiza tehnică. Inițiativa de construire a unor teste noi poate să aparțină fie personalului firmelor de publicare, autorilor de teste și cercetătorilor independenți care își trimit ideile editorilor.

Decizia de a construi un test este luată de obicei atunci când un autor de teste realizează fie că nu există niciun test pentru un anumit scop, fie că testele existente pentru un anumit scop nu sunt adecvate dintr-un motiv sau altul. Considerațiile legate de piața de teste sunt de asemenea centrale în procesul de luare a deciziei în cazul comercializării acestora. În orice caz, din momentul în care s-a luat decizia de a construi un test, scopul și argumentele sale sunt formulate cu atenție în termenii tipurilor de inferențe care pot fi făcute pornind de la scorurile testului, iar autorul trebuie să facă și un plan al testului.

Planificarea unui test presupune specificarea (a) constructelor sau domeniilor de cunoștințe pe care testul le va evalua; (b) tipul de populație pentru care va fi utilizat testul, (c) obiectivele itemilor care vor fi construiți, incluse în cadrul dat de scopul testului și (d) mijloace concrete prin care vor fi colectate și scorate eșantioane de comportament. Acest ultim punct include decizii cu privire la metoda de administrare, formatul stimulilor prezentați de itemi și formatul răspunsurilor, procedurile de scorare ce urmează a fi utilizate. După ce s-au decis aceste aspecte și s-a realizat un plan preliminar al testului, procesul de construire a testului implică de obicei următorii pași:

1. Generarea unui grup de itemi prin scrierea itemilor testului sau crearea lor în alt mod, precum și a procedurilor de administrare și scorare ce urmează a fi utilizate;
2. Grupul de itemi este supus analizei de item calitative realizate de evaluatori și, după caz, itemii sunt revizuiți sau înlocuiți;
3. Itemii generați și revizuiți sunt probați pe eșantioane reprezentative pentru populația căreia îi este destinat testul;
4. Se evaluează administrările de probă ale grupului de itemi prin analiză de itemi cantitativă și analiză calitativă suplimentară.
5. Se adaugă, se șterg și/sau se modifică itemi după nevoie, pe baza analizelor de item cantitative și calitative;

6. Se realizează administrări de probă suplimentare cu scopul de a verifica dacă statisticile itemilor rămân stabile în diferite grupuri — un proces cunoscut drept *validare încrucișată* — până când se obține un set satisfăcător de itemi;
7. Se standardizează sau se fixează lungimea testului și secvențialitatea itemilor, precum și procedurile de administrare și scorare ce urmează a fi utilizate în forma finală a testului pe baza analizelor ulterioare;
8. Se administrează testul unui nou eșantion de indivizi — selecționați cu atenție pentru a reprezenta populația de subiecți căreia îi este destinat testul — pentru a obține date normative sau criterii de performanță, indicatori ai fidelității și validității scorurilor și statistici la nivel de item pentru versiunea finală a testului.
9. Testul se publică în forma sa finală, însoțit de un manual de administrare și scorare și de o documentație ce conține datele de standardizare, studiile de fidelitate și validitate și materialele necesare pentru administrarea și scorarea testului (vezi AERA, APA, & NCME, 1999, cap. 6).

Nu uitați!

Clasificarea testelor în linii mari în teste de „aptitudini" și de „personalitate" este o practică standard. Această distincție tradițională — utilizată în mod repetat în acest capitol pentru a evita dificultățile — se bazează pe ideea că anumite teste sunt construite în principal pentru a evalua aspecte ale comportamentului cognitiv, în timp ce altele sunt construite pentru a evalua aspecte ale comportamentului legate de funcționarea emoțională. Cu toate acestea, atunci când luăm în considerare itemii unui test și testul ca întreg, este important să ne amintim că factorul cognitiv și cel emoțional sunt inseparabili și că eșantioanele de comportament reflectă toate aspectele funcționării unei persoane.

Pentru ca un test să fie comercializat, acești pași pot dura ani și poate fi nevoie să fie repetați de câteva ori dacă rezultatele inițiale nu sunt adecvate. În plus, cele mai multe teste standardizate sunt revizuite periodic datorită învechirii graduale a normelor, criteriilor de performanță și a conținutului testului. Unele instrumente standardizate utilizate în testarea pe scară largă, cum ar fi SAT (cunoscut anterior drept Scholastic Aptitude Test), sunt revizuite și rafinate aproape continuu. În mod evident, testele care sunt utilizate limitat, așa cum se întâmplă în cadrul unei clase de elevi sau în studii specifice de cercetare nu trec printr-un proces atât de riguros. Cu toate acestea, ele trebuie să fie construite cu atenție, conform specificațiilor prestabilite, iar rezultatele lor trebuie să poată fi susținute psihometric — în termenii caracteristicilor itemilor, fidelității și validității lor — pentru a-și putea atinge scopurile cu succes.

Tipuri de itemi

Varietatea itemilor care alcătuiesc testele psihologice este imensă. O astfel de varietate sfidează o clasificare ușoară. Itemii testului, precum și testul ca întreg, pot diferi nu doar în termenii conținutului și formatului, ci și în termenii mediului de administrare, manierei în care sunt scorați și tipului de procesare pe care o cer din partea subiecților testați. Una din distincțiile cele mai elementare privind itemii unui test se referă la tipul de răspunsuri pe care îl cer de la subiecții testați. Din acest punct de vedere, toți itemii testului pot fi clasificați în două mari categorii, și anume itemi cu răspuns la alegere și itemi cu răspuns construit. Testele construite pentru a evalua aptitudinile și cele destinate evaluării personalității utilizează fie unul, fie ambele tipuri de itemi, în

funcție de natura eșantioanelor de comportament necesare pentru scopul testării. În mod similar, ambele tipuri de itemi pot fi utilizate fie în testarea de grup, fie în testarea individuală. Info 6.2 oferă informații despre obținerea unor exemple de diferite tipuri de itemi.

Itemi cu răspuns la alegere

Itemii cu răspuns la alegere, cunoscuți și ca *itemi obiectivi* sau *cu răspuns fix*, au o natură închisă; ei prezintă un număr limitat de alternative din care subiectul testat trebuie să aleagă. În testele de aptitudini, itemii de acest tip includ alegeri multiple, adevărat-fals, ordonarea după rang și gruparea, precum și itemi care cer o rearanjare a opțiunilor oferite. De obicei, itemii obiectivi din testele de aptitudine sunt scorați simplu, de tipul corect–greșit, deși este posibil, de asemenea, să se acorde credite parțiale pentru anumite opțiuni de răspuns. Orice persoană care a fost școlarizată în Statele Unite în ultimele câteva decenii își poate aminti exemple de diferite tipuri ale itemilor de aptitudine cu răspuns la alegere, atât din testele standardizate, cât și din testele realizate de profesori. Itemii cu răspuns la alegere nu au fost întotdeauna utilizați atât de frecvent ca în prezent în Statele Unite, nici nu sunt utilizați la fel de frecvent peste tot în lume. De fapt, mare parte a criticilor testării standardizate din educația din Statele Unite se centrează pe utilizarea pervazivă în cadrul testelor a itemilor cu răspuns la alegere — mai ales a itemilor cu alegeri multiple — și pe slăbiciunea lor percepută din punct de vedere pedagogic. Mulți critici ai „testării standardizate" utilizează acest termen în sens larg, incorect, ca un sinonim pentru testele care au un format de itemi cu alegere multiplă (vezi Mitchell, 1992; Sacks, 1999).

Info 6.2

Cum pot fi găsite exemple de diferite tipuri de itemi

- Site-ul Educational Testing Service Web (http://www.ets.org) oferă legături către câteva dintre programele sale majore de testare. Eșantioane de itemi sunt disponibile în secțiunile de pregătire pentru test ale site-urilor web ale acestor programe. De exemplu:
  - The College Board Web site (http://www.collegeboard.com) oferă miniteste SAT Verbal și Matematică — cu tipul de itemi utilizat în SAT — pe care utilizatorii interesați le pot completa gratuit.
  - Site-ul Graduate Record Examinations (GRE) (http://www.gre.org) oferă o versiune de exersare a Testului General GRE.
  - Întrebări din diferite secțiuni ale Test of English as a Foreign Language (TOEFL) pot fi găsite pe site-ul TOEFL (http://www.toefl.org).
- Exemple de itemi din diferite teste de aptitudine și personalitate sunt disponibile la http://www.schuhfried.co.at, un site pentru Schuhfried Company, o organizație austriacă ce comercializează produse software pentru administrarea computerizată a testelor.
- Imagini ale instrumentelor utilizate în multe teste de sensibilitate și performanță pot fi găsite în catalogul *Evaluation and Assessment* al Lafayette Instrument Company, disponibil la http://www.licmef.com/downloads.htm#cat
- Cataloagele companiilor care publică teste oferă descrieri ale conținutului testelor și deseori includ și exemple de itemi. Cataloagele de pe internet tind să enumere și să descrie testele, dar, de obicei, nu includ exemple de itemi. Utilizatorii interesați pot obține cataloage tipărite contactând editorii (vezi Anexa B).

În testele de personalitate, itemii obiectivi pot fi dihotomici sau politomici. Itemii *dihotomici* cer o alegere între două alternative (ex.: adevărat-fals, da-nu, plăcut-neplăcut etc.) în timp ce itemii *politomici* prezintă subiectului testat trei sau mai multe (de obicei

un număr impar, 3, 5 sau 7) răspunsuri alternative la o afirmație. Aceste alternative sunt de obicei gradate pe o scală în termenii gradului de acceptare (*plăcut, indiferent, neplăcut*), intensității acordului (de la *acord puternic* la *dezacord puternic*), frecvență (e.g. de la *niciodată* la *foarte frecvent*) și așa mai departe — punctul de mijloc semnificând de obicei o poziție neutră, nesigură, de mijloc.

Itemi cu alegere forțată

Itemii obiectivi care cer ca subiectul testat să aleagă una sau mai multe alternative care îl caracterizează cel mai mult sau cel mai puțin se numesc *itemi cu alegere forțată*. Fiecare dintre alternativele dintr-un set de alegeri forțate reprezintă un construct diferit, dar se potrivesc în termenii dezirabilității sociale, astfel încât par la fel de atractive sau la fel de neatractive subiecților testați. Acest tip de item este utilizat în principal în inventarele de personalitate multidimensionale (i.e. inventare construite pentru a evalua câteva constructe ale personalității) pentru a controla tendința subiecților de a răspunde în direcția pe care o percep ca fiind mai dezirabilă social. Totuși, alternativele alegerii forțate sunt deseori împerecheate astfel încât fiecare alegere pe care o face subiectul testat limitează intervalul posibil al scorurilor lor la alt construct sau trăsătura evaluată de testul multidimensional. În acest caz, rezultatele au caracter ipsativ și nu pot fi interpretate într-o manieră normativă. Scorurile *ipsative* sunt în mod esențial numere ordinale care reflectă rangul acordat de subiecți constructelor evaluate pe scalele din cadrul unui test cu format de alegeri forțate. Aceasta înseamnă că mărimea relativă a scorurilor la fiecare dintre scalele unui astfel de test poate fi determinată numai prin compararea cu alte scoruri obținute de același individ la alte scale ale testului, și nu cu scorurile obținute de grupuri normative. Mai mult, formatul de alegeri forțate nu poate elimina în totalitate influența dezirabilității sociale și chiar poate interfera cu relația de testare (vezi Capitolul 7 pentru o definiție a *relației*). În ciuda acestor probleme, itemii cu alegere forțată sunt încă utilizați, mai ales în cazul inventarelor de interese — cum ar fi Myers-Briggs Type Inventory (MBTI) — și în teste al căror scop principal este clasificarea indivizilor în categorii reciproc exclusive. Unele teste cu format de alegere forțată (Jackson Vocational Interest Survey) evită problema caracterului ipsativ prin împerecherea alternativelor extrase din două seturi diferite de scale paralele, astfel încât intervalul scorurilor de pe fiecare scală nu este supus restricțiilor.

Avantajele itemilor cu răspuns la alegere

Itemii obiectivi reprezintă cel mai popular și mai frecvent utilizat tip de itemi ai unui test. Avantajele lor derivă din ușurința și obiectivitatea cu care acești itemi pot fi scorați, ceea ce duce la o economie de timp semnificativă și la sporirea fidelității scorurilor obținute la test; problema erorii de scorare este aproape irelevantă pentru itemii de acest tip, cu excepția greșelilor legate de rutină. Mai mult, itemii cu răspuns la alegere utilizează timpul în mod eficient pentru că mulți dintre ei pot fi administrați într-o perioadă determinată de timp, ceea ce nu se întâmplă cu itemii cu răspuns construit. Deși pot fi administrate și individual, cele mai multe teste care utilizează itemi cu răspuns la alegere sunt destinate testării în grup.

Toate răspunsurile la itemi obiectivi pot fi transformate cu ușurință și fidel pe o scală numerică în scopul scorării, fapt care simplifică în mare măsură analiza cantitativă a acestor itemi. În testele de aptitudini, răspunsurilor corecte și incorecte le sunt atribuite de obicei valori 0 sau, respectiv, 1; ocazional, variații ca

2, 1 sau 0 sunt disponibile pentru notări parțiale. În testele de personalitate, itemii dihotomici sunt, de asemenea, scorați 1 sau 0, dacă răspunsul subiectului testat este sau nu în direcția constructului pe care testul este destinat să îl măsoare. Alternativele prezentate în itemii politomici sau cu răspuns multiplu pot fi transpuse pe diferite scale numerice, cum ar fi 5, 4, 3, 2, 1 sau +2, +1, 0, -1 sau -2 sau pot fi reduse la un format de scorare binară (0 sau 1) prin colapsarea categoriilor.

Dezavantajele itemilor cu răspuns la alegere

În ciuda avantajelor lor, itemii cu răspuns la alegere sunt mai susceptibili decât itemii cu răspuns construit la anumite probleme. În testele de aptitudini, problema majoră legată de itemii obiectivi se situează în jurul ghicirii răspunsurilor. Posibilitatea ghicirii corecte este permanent prezentă atunci când răspunsurile trebuie să fie selectate pur și simplu. În itemii dihotomici, cum ar fi adevărat-fals, probabilitatea de a ghici răspunsul corect este substanțială, 50%. Atunci când subiecții testați ghicesc răspunsurile corecte la itemii obiectivi, gradul de eroare introdus în scorurile lor variază în funcție de factorii (ex.: pură întâmplare, cunoștințe parțiale, formularea itemilor etc.) responsabili pentru ghicirea răspunsului corect. În mod similar, răspunsurile incorecte la itemii obiectivi pot să apară cu ușurință ca rezultat al grabei, neatenției, neglijenței, prefăcătoriei sau orice alt factor întâmplător fără legătură cu nivelul de cunoștințe sau cu aptitudinea subiectului testat în aria acoperită de item.

În testarea personalității, obiectivele intenționate ale itemilor cu răspuns la alegere pot fi subminate cu ușurință dintr-un număr și mai mare de motive. Acestea includ nu numai răspunsul la întâmplare sau neglijent, ci și seturi de răspunsuri ale subiectului

testat care, intenționat sau neintenționat, conduc pe un drum greșit. În funcție de contextul în care are loc testarea și de setul mental specific al subiectului testat, răspunsurile la testele de personalitate pot conduce greșit fie într-o direcție pozitivă, fie în una negativă. De exemplu, bineînțeles că indivizii care răspund la un inventar de personalitate în contextul candidării la un post vor alege să se prezinte pe ei înșiși într-o lumină mai favorabilă decât persoanele testate pentru a stabili dacă o tulburare psihică poate fi utilizată ca factor atenuant în determinarea culpabilității în cazul unui proces de crimă. În mod clar, răspunsurile la itemii obiectivi ai testelor de personalitate pot fi manipulați cu mai mare ușurință de subiecții testați decât răspunsurile la itemii testelor de aptitudini, care nu pot fi falsificate într-o direcție pozitivă, cu excepția trișării (pentru o tratare completă a numeroaselor aspecte ale trișării la teste, vezi Cizec, 1999). Datorită vulnerabilității lor la distorsiune, multe inventare de personalitate utilizează seturi de itemi speciale, scale de validitate sau alte instrumente construite special pentru a detecta răspunsurile care induc în eroare sau cele neglijente.

Toate posibilitățile pe care tocmai le-am subliniat pot diminua fidelitatea și validitatea scorurilor unui test. Deși itemii cu răspuns construit sunt de asemenea susceptibili la unele dintre aceste probleme, ghicirea răspunsului corect la testele de aptitudine cu răspuns construit este mai dificilă și, prin urmare, mai puțin probabilă. Pentru subiecții testați este o provocare mult mai mare să răspundă la testele proiective sau la alte instrumente de evaluare a personalității cu răspuns construit în moduri care induc în eroare. Mai mult, datorită naturii relativ nestructurate a acestor instrumente, chiar și atunci când subiectul testat încearcă în mod conștient să înșele, acestea pot oferi unele informații utile.

Pregătirea itemilor cu răspuns la alegere este o sarcină dificilă, care consumă timp și cere o elaborare specializată a testului și aptitudini de scriere a itemilor, pe lângă o familiaritate mare cu constructul sau disciplina de care se ocupă testul. Itemii obiectivi slab pregătiți pot oferi neintenționat indicii subiecților testați sau pot fi formulați în termeni care sunt în beneficiul sau în detrimentul subiecților testați. În special itemii cu alegere multiplă scrișI cu neglijență includ deseori alternative care sunt (a) incompatibile gramatical cu prima parte a itemului, (b) susceptibile la diferite interpretări din cauza impreciziei formulării în cuvinte sau (c) atât de nerezonabile, încât pot fi eliminate cu ușurință.

În cele din urmă, itemii cu răspuns la alegere sunt în mod clar mai puțin flexibili decât itemii cu răspuns construit în raport cu intervalul posibil de răspunsuri. Prin urmare, ei nu oferă nicio oportunitate pentru a evalua caracteristicile speciale sau unice ale unui anumit subiect sau cele aflate în afara spectrului de alternative oferit.

Itemii cu răspuns construit

Caracteristica esențială a *itemilor cu răspuns construit*, cunoscuți și ca *itemi cu răspuns liber* este că ei sunt itemi cu final deschis. Varietatea lor este nelimitată, deoarece răspunsurile construite pot presupune răspunsuri în scris sau răspunsuri orale libere, performanțe de orice fel și orice tip de produs.

În testele de aptitudine, cel mai obișnuit tip de itemi cu răspuns construit sunt sarcinile care cer scrierea unui eseu și completarea-spațiilor-libere. Singurele constrângeri aplicate itemilor cu răspuns liber din testele psihologice sunt condițiile impuse de instrucțiunile testului. Sunt indispensabile instrucțiuni complete

și reguli procedurale pentru administrarea standardizată și scorarea tuturor testelor, inclusiv a testelor cu răspuns liber. Indicațiile pentru administrarea testelor cu răspuns construit trebuie să includă stipulații cu privire la aspecte cum ar fi (a) limita de timp; (b) mediul de răspuns, maniera sau lungimea răspunsului cerut; și (c) dacă se permite accesul la materiale sau instrumente relevante pentru test (manuale, calculatoare, computere etc.)

Interviurile, chestionarele de date biografice și observațiile comportamentale sunt instrumente pentru evaluarea personalității care se bazează deseori pe răspunsuri la itemi cu final deschis. În testarea personalității, utilizarea răspunsurilor construite este limitată în principal la *tehnicile proiective*. Aceste metode cer de obicei subiecților să răspundă la stimuli ambigui prezentați sub forma imaginilor (inclusiv pete de cerneală) sau a materialelor verbale, cum ar fi cuvinte sau fraze incomplete. Unele tehnici proiective cer autoexprimarea prin desene sau alte tipuri de acțiuni. Ideea de bază în toate aceste metode — care își au originea în cadrul clinic și sunt utilizate în principal în acest cadru — este de a le prezenta subiecților sarcini cu un grad minim de structură, astfel încât ei să poată răspunde cât mai liber posibil și, în acest proces, să reveleze aspecte semnificative ale personalității lor. În contrast cu inventarele, anchetele și alte astfel de instrumente obiective construite pentru a evalua constructe specifice sau constelații de trăsături de personalitate, tehnicile proiective oferă o abordare a evaluării mai puțin concentrată, mai globală. În general, avantajele și dezavantajele itemilor cu răspuns construit sunt opuse celor prezentate de itemii cu răspuns la alegere. Dar, fără îndoială, acestea merită să fie menționate.

Avantajele itemilor cu răspuns construit

Chiar și atunci când nu sunt administrați individual, itemii cu răspuns construit oferă eșantioane mai bogate din comportamentul indivizilor examinați și permit caracteristicilor unice ale acestora să iasă la suprafață. Itemii cu final deschis oferă o largă gamă de posibilități și abordări mai creative ale testării și evaluării decât itemii cu răspuns la alegere. Mai mult, sarcinile cu răspuns construit aduc la suprafață eșantioane autentice de comportament al subiecților în domenii specifice, în contrast cu alegerea dintre alternative prestabilite. Dacă se urmărește evaluarea abilităților de scriere, a memoriei, cunoștințelor matematice, a abilităților mecanice, a capacității de a conduce sau orice alt tip de performanță, eșantioane actuale pentru ceea ce poate face un individ reprezintă singurul standard indiscutabil.

Dezavantajele itemilor cu răspuns construit

Dezavantajul major al itemilor cu răspuns construit este legat de fidelitatea scorurilor și, în consecință, și de validitate (vezi secțiunea despre relația dintre fidelitate și validitate din Capitolul 4). Aceste dezavantaje apar din modul în care sunt scorate răspunsurile construite și din limitele practice pe care răspunsurile de acest tip le impun lungimii testului.

Scorarea răspunsurilor construite, atât în cazul testelor de aptitudine, cât și al celor de personalitate, cere întotdeauna mai mult timp și și este mai complexă decât scorarea răspunsurilor la alegere pentru că, invariabil, intervine un anumit grad de subiectivitate. Chiar și atunci când *rubricile de scorare* (instrucțiuni care specifică criteriile, principiile și regulile care urmează a fi utilizate în scorare și care oferă exemple ilustrative) sunt examinate și aplicate cu atenție, există întotdeauna posibilitatea ca

un răspuns să fie evaluat în mod diferit de evaluatori diferiți datorită unicității lui sau altor factori. Evaluarea fidelității scorurilor atribuite de diferiți evaluatori este un aspect indispensabil și costisitor al utilizării testelor alcătuite din itemi cu răspuns construit. Deși diferențele inter-evaluatori nu pot fi eliminate complet, ele pot fi minimizate cu siguranță prin intermediul procedurilor de scorare explicite, testate într-un studiu-pilot, precum și prin instruirea adecvată a evaluatorilor.

Scorarea răspunsurilor construite obținute prin metode proiective utilizate în evaluarea personalității reprezintă o provocare specială prin faptul că subiectivitatea evaluatorului poate intra în joc în mai multe moduri decât în scorarea răspunsurilor construite date la testele de aptitudine. Mai mult, tehnicile proiective se adaptează mai ușor la utilizarea metodelor de administrare și scorare informale și deseori idiosincratice, care pot slăbi integritatea lor psihometrică (vezi Lanyon & Goodstein, 1997, cap. 4).

Lungimea testului este un alt factor care afectează fidelitatea scorurilor la testele ce utilizează răspunsul construit. Deoarece aceste răspunsuri necesită mai mult timp pentru administrarea și scorarea unui test, numărul de itemi care pot fi incluși în testele cu răspuns construit este de obicei mult mai mic decât în testele cu răspuns la alegere. Așa cum s-a discutat în Capitolul 4, dacă toate celelalte aspecte rămân constante, testele mai scurte sunt mai predispuse la erori ale eșantionării conținutului și produc scoruri care sunt mai puțin consistente decât testele mai lungi. Astfel, din punctul de vedere al consistenței interne, scorurile obținute la testele cu răspuns construit tind să aibă fidelitate mai mică decât testele cu răspuns la alegere.

O complicație suplimentară a itemilor cu răspuns construit privește aspectul *lungimii răspunsului*. Din moment ce răspunsurile

mai lungi conţin mai mult material decât cele mai scurte, variaţiile lungimii răspunsurilor construite pot influenţa scorurile în mod considerabil. Acest fapt este relevant în special pentru tehnicile proiective, deoarece este probabil ca răspunsuri proiective mai lungi — sau mai elaborate — să conţină mai multe elemente care pot fi scorate (adică semnificative din punct de vedere psihologic) decât cele mai scurte. Mai mult, instrumentele proiective care permit ca numărul şi lungimea răspunsurilor să varieze prezintă încă un factor care complică investigarea proprietăţilor psihometrice datorită lipsei de uniformitate între subiecţii testaţi. Testul Rorschach este un exemplu proeminent de astfel de problemă, aşa cum dovedeşte controversa de lungă durată privind impactul productivităţii răspunsurilor asupra scorării şi interpretării Rorschach (vezi Groth-Marnat, 1997, p. 399; Meyer, 1992).

Analiza de item

În ultimele câteva decenii, domeniul construirii testelor şi tehnicile de analiză de item au traversat o etapă de tranziţie care a început să aibă efecte în mod gradual, dar care modifică fundamental natura testelor psihologice. Această tranziţie se datorează în parte uşurinţei şi eficienţei cu care datele unui test pot fi colectate, arhivate, analizate, recuperate şi răspândite cu ajutorul computerelor. În plus, din anii '60, metodologia oferită de noile abordări ale construirii testelor psihologice cunoscută drept teoria răspunsului la item (TRI) sau teoria trăsăturii latente a suplimentat în mod constant — şi în unele cazuri a înlocuit — metodele tradiţionale de construire a testelor. Deşi metodele TRI pot fi şi sunt utilizate în construirea testelor creion-hârtie şi a testelor de

lungime fixă bazate pe computer, cel mai proeminent avantaj faţă de metodologia tradiţională este faptul că ele permit un format de test mai flexibil şi mai eficient prin testarea adaptativă computerizată (TAC). În TAC, secvenţele de itemi pot fi ajustate individual la nivelele de aptitudine ale subiecţilor testaţi sau la poziţia atinsă de subiecţi pe baza răspunsurilor anterioare la oricare trăsătură pe care testul este destinat să o evalueze. În următoarea secţiune a acestui capitol, procedurile tradiţionale de analiză de itemi sunt mai întâi prezentate, apoi sunt urmate de o discutare a metodologiei TRI.

Metodele cantitative ale analizei de item

Pentru testele psihologice în general cel mai important aspect al analizei cantitative de item se centrează pe mărimile statistice care se referă la *validitatea itemului*. Întrebarea la care încearcă să răspundă indicii de validitate este dacă un item specific contribuie la un test extrăgând informaţii care determină un progres din perspectiva scopului testului. De obicei, psihometricienii se referă la mărimile statistice ale validităţii itemului ca indici ai *capacităţii de discriminare a itemului*, deoarece rolul lor este de a dezvălui măsura în care un

ATENŢIE!

Termenul *discriminare* a dobândit o conotaţie negativă în limbajul comun din cauza frecventei sale asocieri cu tratamentul incorect al femeilor şi grupurilor care aparţin minorităţilor rasiale.

Dimpotrivă, în domeniul psihometriei, *discriminarea* este considerată o trăsătură dezirabilă a itemilor testului. Ea se referă la măsura în care itemii produc răspunsuri care diferenţiază în mod precis între subiecţii testaţi la nivelul dimensiunii pe care testul este construit să o evalueze.

item diferențiază în mod precis subiecții testați în raport cu trăsăturile sau comportamentele pe care testul este destinat să le măsoare. În special în cazul testelor de aptitudine, analiza de item include proceduri destinate să determine două caracteristici suplimentare ale itemilor care le influențează validitatea, și anume *dificultatea itemului* și *corectitudinea itemului*. Toate aceste caracteristici ale itemului pot fi evaluate calitativ și cantitativ. Evaluarea calitativă este desfășurată de experți în domeniul căruia îi este destinat testul, aceștia examinând conținutul itemilor pentru a stabili adecvarea acestora și nivelul de dificultate, precum și pentru a stabili dacă itemii reflectă obiectivele specificate ale testului. Conținutul itemilor este de asemenea examinat din punctul de vedere al posibilei sale incorectitudini sau al caracterului ofensator pentru orice grup de subiecți potențiali. Evaluarea cantitativă a dificultății și capacității de discriminare a itemului este realizată prin statistici care stabilesc dacă itemii funcționează așa cum s-a prevăzut atunci când sunt administrați tipului de subiecți cărora le este destinat testul.

Dificultatea itemului

Rolul dificultății itemului în testarea aptitudinilor

Dată fiind afirmația evidentă că nivelul de dificultate al unui test ca întreg depinde de nivelele de dificultate ale itemilor individuali care îl alcătuiesc, rezultă că un test ușor este alcătuit din itemi ușori și un test dificil este alcătuit din itemi dificili. Această premisă aparent simplă devine ceva mai complicată îndată ce luăm în considerare faptul că dificultatea este o chestiune relativă. Cât de dificil este itemul unui test depinde nu numai de simplitatea sa intrinsecă sau accesibilitatea sa, ci și de nivelul aptitudinii subiectului testat. De exemplu, utilizarea adecvată a verbului *être* (a fi) — cel mai comun verb din limba franceză — este o sarcină mult mai ușoară pentru un student dintr-un curs avansat de limbă franceză, decât pentru unul care începe să studieze această limbă. Astfel, pentru a calibra în mod adecvat nivelul de dificultate al unui test, este nevoie de indicatori ai dificultății *relative* a itemilor pentru unul sau mai multe grupuri relevante. Autorii de teste utilizează acești indicatori pentru a determina adecvarea itemilor la populație și la scopul pentru care este construit testul, precum și pentru a decide unde să fie plasați itemii în cadrul unui test.

Nu uitați!

Așa cum cadrele de referință pentru interpretarea scorurilor unui test discutate în Capitolul 3 pot fi normative sau legate de criteriu, tot așa și dificultatea itemilor unui test poate fi determinată fie pe o bază absolută, fie pe o bază normativă. Ambele aspecte trebuie luate în considerare în procesul de construire și evoluție a unui test, ținând cont în mod specific de populația de subiecți urmărită și de scopul testului.

Cum se stabilește dificultatea unui item?

În timpul stadiilor inițiale ale construirii testului, atunci când este generat grupul de itemi, autorii testului pot stabili dificultatea itemilor pe care îi creează pe baza unor standarde mai mult sau mai puțin obiective extrase din specificațiile făcute pentru test, pe baza criteriilor agreate de experții din domeniul căruia i se adresează testul sau a aptitudinilor cognitive acoperite de test. De exemplu, un standard care poate fi aplicat pentru a calibra dificultatea cuvintelor este frecvența cu care acestea sunt utilizate în limbaj. Astfel, într-un test de vocabular, itemii ușori sunt

cuvintele utilizate frecvent de către vorbitorii limbii respective, în timp ce itemii cei mai dificili sunt alcătuiți din cuvinte care apar rar și ar putea fi nefamiliare pentru cei mai mulți subiecți testați. În mod similar, într-un test de aritmetică, itemii individuali pot fi echilibrați în termenii dificultății lor pe baza complexității evidente a operațiilor pe care le necesită, cum ar fi înmulțirea numerelor întregi versus înmulțirea fracțiilor și așa mai departe.

Odată ce un set de itemi este administrat unuia sau mai multor grupuri, pot fi obținuți indicatorii cantitativi ai dificultății itemului, care abordează acest aspect dintr-o perspectivă normativă. În analiza itemilor testului din perspectivă normativă, elementul esențial de informație utilizat pentru a determina dificultatea itemului este procentul de subiecți testați care răspund corect la un item, cunoscut și ca *proporție* (sau procent) *de răspunsuri corecte*, sau, pe scurt, *p*. Cu cât este mai mare procentul de răspunsuri corecte, cu atât este mai ușor itemul testului. Din moment ce valorile *p* ale itemilor depind în întregime de nivelul aptitudinii grupurilor cărora li s-au administrat itemii, alcătuirea acestor grupuri este destul de importantă și trebuie să reflecte componența populației căreia i se adresează testul.

Procentul de răspunsuri corecte sau valorile *p* sunt numere ordinale care, asemenea rangurilor percentilice, nu reprezintă unități egale. Din acest motiv, presupunând că trăsătura pe care o măsoară un item are o distribuție normală, valorile *p* sunt deseori transformate în valori *z* utilizând Tabelul ariilor situate sub curba normală (vezi Anexa C). Odată ce valorile *p* sunt transformate în valori *z*, nivelele de dificultate relativă a itemilor pot fi comparate între grupuri diferite prin administrarea unui set comun de itemi — numiți itemi *ancoră* — în două sau mai multe grupuri. Formulele de estimare a nivelelor de dificultate

ale itemilor suplimentari în grupurile în chestiune pot fi obținute pe baza relațiilor stabilite între itemii ancoră. Această metodă, cunoscută drept *scalare absolută*, a fost elaborată de Thurstone (1925). Ea permite ca dificultatea itemilor să fie plasată pe o scală numerică uniformă pentru eșantioane de subiecți aflați la diferite nivele de aptitudine, cum ar fi studenții din diferite clase. Info 6.3 prezintă un exemplu numeric simplu al modului în care se desfășoară această metodă utilizând rezultatele obținute la cinci itemi administrați în două grupuri. Deoarece toți cei cinci itemi din exemplu au valori *p* mai mari (și valori *z* mai mici) în grupul B decât în Grupul A, putem presupune că Grupul B funcționează la un nivel mai avansat decât Grupul A în aria de conținut sau aptitudine pe care acești itemi o vizează. Figura 6.1 ilustrează dificultatea relativă a celor cinci itemi pentru cele două grupuri și demonstrează grafic că nivelele de dificultate ale celor cinci itemi pentru cele două grupuri corelează puternic și pozitiv. Tipul de date prezentate în Info 6.3 și Figura 6.1 poate fi utilizat pentru a estima nivele de dificultate ale itemilor suplimentari pentru un grup pe baza valorilor dificultății obținute pentru celălalt grup prin intermediul analizei regresiei (vezi capitolul 5). Aceste tipuri de proceduri se aplică în echivalarea testelor și a scorurilor obținute la teste pe baza testelor ancoră și a grupurilor fixe de referință (vezi Capitolul 3).

Nivele de dificultate a itemului, nivele de dificultate a testului și scopul testului

Pentru orice grup dat de subiecți, scorul mediu la un test este același cu dificultatea medie a itemilor săi. Astfel, dacă procentul *mediu* de răspunsuri corecte (p) pentru itemii unui test este 80%, scorul mediu obținut la test va fi tot 80%. Semnificația relației

Info 6.3

Transformarea dificultății itemului din proporția de răspunsuri corecte (*p*) în unități ale curbei normale (*z*)

Dificultatea itemului poate fi reprezentată în unități ale curbei normale (valori *z*) dat fiind că se presupune că trăsătura măsurată de un item are o distribuție normală.

Valoarea *z* pentru un item este obținută prin localizarea proporției de răspunsuri corecte (i.e. valoarea *p*) în Tabelul ariilor situate sub curba normală (vezi Anexa C): *valorile* p *mai mari de 0,50* se află în coloana 3 a tabelului și li se atribuie valorile *z* corespunzătoare cu un semn *negativ; valorile p sub 0,50* se află în coloana 4 a tabelului și valorilor *z* corespunzătoare li se atribuie un semn *pozitiv*. Dacă $p = 0,50$, valoarea *z* pentru acest item este zero.

Exemplu numeric pentru cinci itemi administrați în două grupuri:

Numărul itemului	Valoarea *p* în grupul A[a]	Valoarea *z* în Grupul A[b]	Valoarea *p* în Grupul B[a]	Valoarea *z* în Grupul B[b]
1	0,841	-1,00	0,894	-1,25
2	0,50	0,00	0,691	-0,50
3	0,067	+1,50	0,159	+1,00
4	0,023	+2,00	0,067	+1,50
5	0,977	-2,00	0,994	-2,51

[a] Valorile *p* reprezintă proporția de indivizi din Grupul A și din Grupul B care au răspuns corect la item.

[b] Valorile *z* pentru cei mai ușori itemi sunt mari și negative, în timp ce valorile pentru cei mai dificili itemi sunt mari și pozitive.

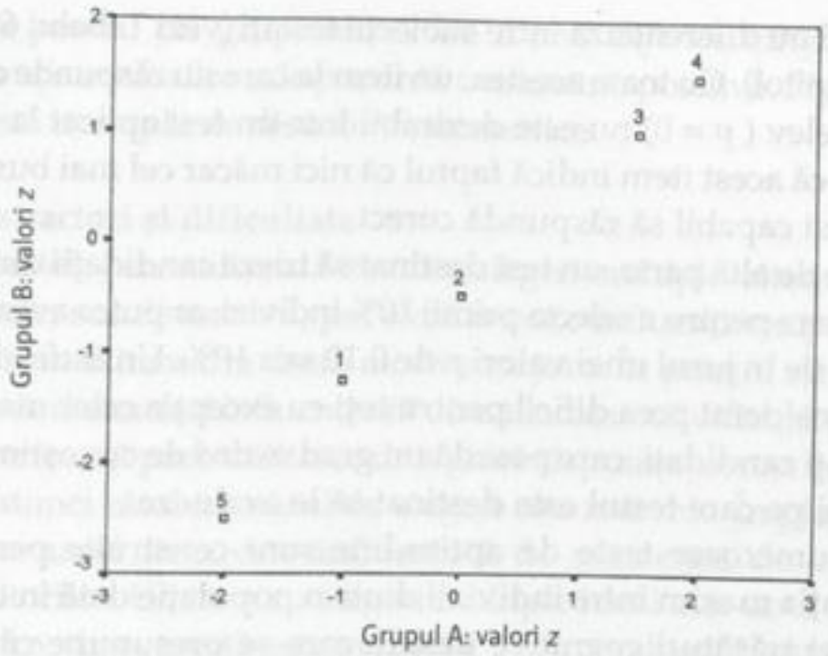

Figura 6.1. Diagrama de împăștiere a dificultății relative a celor cinci itemi pentru Grupurile A și B (vezi Info 6.3)

dintre dificultatea itemului, scopul testului și nivelul de aptitudine al populației de subiecți testați pentru care este construit testul poate fi clarificată prin câteva exemple.

- Testele de achiziții școlare sunt construite pentru a evalua măsura în care elevii dintr-o clasă stăpânesc materia unei discipline școlare. În cele mai multe cadre academice, o notă în intervalul 70%–79% este considerată a fi medie. Pentru a obține această medie, cei mai mulți dintre itemii din testele utilizate la clasă trebuie să poată fi rezolvați de majoritatea elevilor, dat fiind că aceștia stăpânesc conținutul materiei la un nivel pe care profesorul îl consideră mediu. Astfel de teste pot include unii itemi — care presupun concepte ce au fost subliniate în timpul predării — la care întreaga clasă să răspundă corect ($p = 1,00$), chiar dacă astfel

de itemi nu diferențiază între subiecții testați (vezi Tabelul 6.1 din acest capitol). Cu toate acestea, un item la care nu răspunde corect niciun elev ($p = 0$) nu este dezirabil într-un test aplicat la clasă, pentru că acest item indică faptul că nici măcar cel mai bun elev nu a fost capabil să răspundă corect.

- Pe de altă parte, un test destinat să trieze candidații dintr-un grup mare pentru a selecta primii 10% indivizi, ar putea avea itemi care să fie în jurul unei valori p de 0,10 sau 10%. Un astfel de test ar fi considerat prea dificil pentru toți cu excepția celor mai înalt calificați candidați, care posedă un grad extins de cunoștințe sau abilități pe care testul este destinat să le evalueze.
- Numeroase teste de aptitudine sunt construite pentru a diferenția maxim între indivizii dintr-o populație dată în termenii unei trăsături cognitive despre care se presupune că are o distribuție normală, cum ar fi inteligența generală sau abilități verbale. În astfel de teste, intervalul de dificultate a itemilor trebuie să fie suficient de larg pentru a include cei mai capabili și cei mai puțin capabili indivizi dintr-o populație potențială de indivizi testați, iar valoarea p a itemilor ar trebui să se situeze în jurul lui 0,50 (sau 50%) pentru a oferi o diferențiere maximă între subiecții testați. Itemii cu valori p extreme (apropiate de 0 sau de 1,00) trebuie evitați în astfel de teste pentru că nu reușesc să diferențieze între subiecții testați și, prin urmare, sunt elemente în plus. Mai mult, dacă indivizii care aparțin populației pentru care este construit un test de o astfel de natură sunt capabili să răspundă corect la toți itemii sau să nu răspundă corect la niciun item, scorurile lor sunt nedeterminate. Așa cum am discutat în Capitolul 3, atunci când itemii unui test sunt prea ușori pentru un anumit grup, se spune că testul are un *plafon insuficient* și distribuția scorurilor sale prezintă o asimetrie negativă; atunci când itemii sunt prea

dificili pentru un grup, testul are o *prag inferior insuficient* și distribuția scorurilor sale prezintă o asimetrie pozitivă. Figura 2.4 prezintă exemple de distribuții asimetrice.

Distractori și dificultate

În testele care utilizează itemi cu alegeri multiple, alternativele incorecte, sau *distractorii*, pot avea o mare influență asupra dificultății unui item din două motive majore. În primul rând, numărul de distractori influențează direct indicatorii dificultății unui item pentru că probabilitatea de a ghici răspunsul corect este mai mare atunci când numărul de alegeri este mai mic. În plus, dificultatea unui item este influențată de asemenea de calibrul distractorilor. Un item cu alegeri multiple ideal este unul în care (a) alternativa corectă este evidentă pentru subiecții testați care cunosc răspunsul și (b) distractorii par la fel de plauzibili pentru cei care nu îl cunosc. Astfel de itemi sunt greu de construit. În măsura în care distractorii par în mod evident răspunsuri greșite, au deficiențe în formularea verbală sau sunt mult mai scurți sau mult mai lungi decât alternativa corectă, ei oferă indicii pe care examinații care nu cunosc răspunsul le pot utiliza pentru a-și restrânge aria de alegere și a selecta răspunsul corect. Pentru a evita aceste probleme și altele în construirea itemilor cu alegeri multiple, autorii testelor ar trebui să urmeze regulile de scriere a itemilor oferite în manuale cum este cel scris de Haladyna (1999). După ce un test este administrat, ar trebui realizată o analiză a distractorilor. Astfel de analize calculează numărul de subiecți testați care au ales fiecare distractor. Examinarea atentă a frecvențelor cu care diferiți distractori sunt aleși de subiecții testați aflați la diferite nivele de aptitudine servește pentru a detecta posibilele scăpări ale itemilor. Dacă un test se află încă în construcție,

distractorii care nu funcționează în mod adecvat (cei care nu sunt aleși de nimeni sau cei aleși cel mai frecvent de subiecții testați situați la nivele înalte de aptitudine) trebuie înlăturați și înlocuiți.

Dificultatea unui item este un concept relevant în testarea personalității?

Este posibil ca sarcinile care alcătuiesc testele de personalitate să nu fie construite pentru a evalua funcționarea cognitivă, dar ele presupun procese cognitive. În instrumentele cu răspuns la alegere, cum ar fi inventarele și chestionarele de personalitate, procesele cognitive relevante sunt legate de aptitudinea subiectului de a înțelege itemii. Prin urmare, nivelul vocabularului și abilitățile de lectură ale potențialilor subiecți supuși testării trebuie să fie luate în considerare în construirea acestor itemi. Pe de altă parte, sarcinile proiective presupun un anumit grad de competență în modul în care urmează a fi exprimate răspunsurile. Cele mai multe instrumente proiective necesită existența unui nivel minim al abilității de exprimare verbală, exprimare prin desen sau alt tip de activitate. Astfel, dificultatea relativă sau ușurința sarcinilor proiective pentru diferite tipuri de examinați trebuie luată de asemenea în considerare în construirea, administrarea și interpretarea acestor instrumente.

Capacitatea de discriminare a itemului

Capacitatea de discriminare a itemului se referă la măsura în care itemii evocă răspunsuri care diferențiază cu precizie între subiecții testați în termenii comportamentului, cunoștințelor sau ai altor caracteristici pe care un test — sau un subtest — este

construit să îl evalueze. Pentru marea majoritate a testelor, puterea de discriminare este calitatea fundamentală pe care itemii trebuie să o dețină pentru a fi incluși în test. În procesul de construire a testului, indicatorii de discriminare a itemului — cunoscuți și ca indicatori ai validității itemului — sunt obținuți utilizând anumite criterii sau un anumit indicator al poziției subiectului testat la constructul evaluat de test. Criteriile utilizate în acest scop pot fi (a) criterii interne care se referă la testul aflat în construcție (i.e scorul total la test), (b) criterii externe de același tip ca cele utilizate pentru a valida testele ca întreg și descrise în Capitolul 5 (vârstă, educație, apartenență la grupuri contrastante ocupațional sau diagnostice etc.) sau (c) combinații ale criteriilor interne și externe.

Criterii de validare a unui item

Alegerea criteriilor în raport cu care sunt validați itemii testului depinde de scopul testului. Testele de aptitudine necesită criterii legate de ariile de conținut sau aptitudinile pe care le evaluează; testele de personalitate necesită criterii relevante pentru trăsături sau aspecte ale comportamentului de care se ocupă testul. Calitatea și adecvarea criteriilor utilizate în validarea itemilor testului au consecințe importante cu privire la selectarea itemilor care vor fi păstrați într-un test și, în consecință, asupra fidelității și validității scorurilor testului.

Atunci când sunt utilizate criterii externe testului pentru a valida itemii, crește validitatea scorurilor testului ca întreg; atunci când este utilizat criteriul intern al testului total pentru a valida itemii, omogenitatea testului crește și, ca urmare, cresc indicii de fidelitate bazați pe consistența inter-itemi. În construirea testelor care evaluează o singură trăsătură unidimensională, cum ar fi vocabularul sau depresia, pentru a valida itemii poate fi folosit

scorul total. Această practică se bazează pe presupunerea că toți itemii din cadrul unor astfel de teste trebuie să coreleze puternic cu scorul total la test și să coreleze și unul cu celălalt. Pe de altă parte, în construirea testelor care evaluează constructe complexe și multifațetate cum ar fi inteligența, itemii se validează în raport cu un criteriu extern care este, de asemenea, mai global. Deoarece itemii acestor teste pot evalua diferite aspecte ale unui construct complex, ei nu trebuie în mod necesar să coreleze puternic unul cu altul și gradul lor de corelație cu scorul total poate varia. De exemplu, cele mai multe scale de inteligență includ un amestec de itemi care se referă la diferite tipuri de abilități asociate cu acest construct — cum ar fi aptitudini verbale, numerice, spațiale și de raționament logic — și oferă scoruri compuse care încorporează performanța la toate tipurile de itemi și sunt validate în raport cu un criteriu extern, cum ar fi achizițiile educaționale. În instrumente de acest tip, itemii sunt grupați de obicei în subteste cu conținut omogen care sunt scorate separat (vezi Tabelul 4.1 și Fig.4.1 din Capitolul 4).

Așa cum am observat, deși validitatea externă și consistența internă sunt obiective dezirabile în construirea unui test, natura constructelor evaluate de test poate să nu permită ca ambele obiective să fie realizate concomitent. Pe lângă limitele impuse de scopul testului, și validarea externă a itemilor testului poate fi nepractică datorită indisponibilității sau inaccesibilității datelor criteriului extern. Un exemplu tipic de astfel de situație este oferit de itemii din testele pregătite de profesor pentru a fi utilizate la clasă, cum sunt cei prezentați în Tabelul 6.1. Atunci când se realizează analiza de item a acestor teste, profesorii nu pot utiliza alt criteriu decât scorul total la test, pentru că, altfel, nu ar fi corect. Testele utilizate la clasă sunt destinate evaluării măiestriei și

Tabelul 6.1 Date ale analizei de item pentru un test administrat la clasă

Numărul itemului	Procentul de răspunsuri corecte (valoare *p*)			Indice *D* (superior-inferior)	Corelație punct biserial (r_{pb})[b]
	Total grup	Grupul superior[a]	Grupul inferior[a]		
1	100%	100%	100%	0	0,00
2	88%	100%	50%	50	0,67
3	38%	100%	0%	100	0,63
4	75%	50%	50%	0	0,13
5	75%	50%	100%	-50	-0,32
6	13%	50%	0%	50	0,43

[a] Grupurile-criteriu superior și inferior sunt alcătuite din studenți ale căror scoruri la întreg testul s-au aflat în 27% din partea superioară, respectiv inferioară a distribuției scorurilor.

[b] Indicele punct biserial este un indicator al corelației dintre performanța fiecăruia dintre subiecții testați la itemii scorați dihotomic (corect-incorect) și scorurile totale obținute la test.

gradului de stăpânire a materiei predate, iar scorurile lor nu trebuie să fie legate de alt factor decât capacitatea elevilor de a stăpâni obiectivele specificate.

Mărimi statistice ale capacității de discriminare a unui item

Toate procedurile statistice utilizate pentru a determina gradul în care itemii discriminează în termenii unui criteriu necesită informații despre (a) performanța la itemi și (b) poziția în raport

cu criteriul pentru indivizii din eşantionul din care sunt extrase mărimile statistice ale capacităţii de discriminare a unui item. Mărimile statistice utilizate în mod tradiţional în acest scop sunt de două tipuri: indicele statistic de discriminare (sau *D*) şi o varietate de indicatori ai corelaţiei.

Indicele capacităţii de discriminare a itemului (*D*) este utilizat în principal pentru teste scorate corect-incorect, dar poate fi aplicat şi în analiza de item a altor teste care utilizează scorarea binară. Pentru a calcula *D*, subiecţii testaţi trebuie clasificaţi în grupuri criteriu distincte fie pe baza scorurilor lor totale la test, fie pe baza unui anumit indicator extern al poziţiei în raport cu criteriul măsurat de test. De obicei se creează grupuri criteriu prin separarea subiecţilor testaţi utilizaţi pentru analiza validităţii itemilor în două grupuri extreme, de exemplu cei care se află în prima şi ultima treime a valorilor criteriului. Fiind create grupurile criteriu superior şi inferior, se calculează procentul de indivizi (*p*) din fiecare grup care răspund corect la item — sau care răspund în direcţia considerată relevantă pentru constructul măsurat de test. *Indicele de discriminare* reprezintă pur şi simplu diferenţa de procent sau de proporţie a subiecţilor din grupurile superior şi inferior al criteriului care au răspuns corect la un item sau au dat un răspuns în direcţia indicată; *D* poate avea valori între +100 şi -100 (sau între +1,00 şi -1,00). Pentru testele de capacitate, indici ai capacităţii de discriminare pozitivi indică faptul că mai mulţi indivizi din grupul-criteriu superior au răspuns corect la un item decât din grupul inferior şi cea mai dezirabilă valoare pentru *D* este în apropierea lui 1,00. Valorile *D* negative indică faptul că itemii în chestiune discriminează în direcţia opusă şi că trebuie modificaţi sau înlăturaţi. Tabelul 6.1 prezintă indicii capacităţii de discriminare pentru şase itemi ai unui test administrat la un

curs de testare psihologică. La itemul 1, cel mai uşor dintre cei şase, au răspuns corect toţi studenţii (*p* = 100%) iar la Itemul 6, cel mai dificil, au răspuns corect numai 13%. Itemul 3, la care au răspuns corect 38% dintre studenţi, a fost relativ dificil şi a fost itemul care a discriminat cel mai mult între ei, cu o valoare *D* de 100. Itemii 4 şi 5 au fost relativ uşori (*p* = 75%), dar de o valoare îndoielnică. Itemul 4 nu a discriminat între cele două grupuri-criteriu extreme, iar itemul 5 a trebuit să fie înlăturat deoarece indicele *D* de -50 indica faptul că itemul discrimina în direcţia greşită.

De asemenea, *coeficienţi de corelaţie de diferite tipuri* pot exprima relaţia dintre performanţa la un item şi poziţia în raport cu criteriul, oferind astfel indicatori ai capacităţii de discriminare a itemului. Tipul de coeficient de corelaţie ales pentru a calcula aceşti indicatori depinde de natura celor două variabile care urmează să fie corelate, scorurile la item şi mărimea la criteriu. De exemplu, atunci când scorurile la un item sunt dihotomice (corect–incorect) şi dimensiunea criteriului este continuă (scorul total la test), coeficientul

Nu uitaţi!

Cei mai mulţi indici ai capacităţii de discriminare a unui item tind să favorizeze itemi cu o dificultate intermediară. De exemplu, dacă procentul de răspunsuri corecte (valoare *p*) la un item pentru eşantionul total este extrem (100% şi 0%), nu va exista nicio diferenţă între valorile *p* ale grupurilor criteriu superior şi inferior pentru acel item, iar indicele *D* este 0. Pe de altă parte, atunci când valoarea *p* pentru grupul total este 50%, este posibil ca indicele *D* să atingă valoarea sa maximă de +100, dacă toţi cei din grupul-criteriu superior răspund corect şi niciunul din grupul criteriu inferior nu răspunde corect. Prin urmare, în toate testele al căror obiectiv este stabilirea diferenţelor dintre indivizi cu privire la o anumită aptitudine sunt preferaţi itemii care se centrează în jurul unui nivel de dificultate de 50%.

de corelație punct biserial (r_{pb}) este utilizat cel mai frecvent. Pe de altă parte, atunci când scorurile la item și dimensiunile criteriului sunt ambele dihotomice, este utilizat coeficientul phi (ϕ). Atât coeficientul de corelație punct biserial, cât și coeficientul phi se întind între -1,00 și +1,00 și se interpretează în același mod ca Pearson *r*. Formulele de calcul pentru acești coeficienți și pentru alte tipuri de coeficienți de corelație utilizați în analiza discriminării unui item pot fi găsite în cele mai multe manuale de statistică. Oricum, corelațiile pozitive înalte indică o relație directă și puternică între item și criteriu, iar corelațiile negative înalte indică o relație inversă și puternică între item și criteriu, iar corelațiile scăzute indică o relație slabă între cele două. Tabelul 6.1 enumeră și corelațiile punct biseriale între itemi, și scorurile la test pentru fiecare dintre cei șase itemi prezentați mai sus.

O observație cu privire la viteză

Ori de câte ori testele de aptitudine au limite de timp, viteza realizării sarcinii influențează într-o anumită măsură scorurile obținute la test. Acest subiect a fost discutat în Capitolul 4, în legătură cu problemele ridicate de testele de rapiditate în calcularea coeficienților de fidelitate prin înjumătățire și trebuie să fie luate în considerare în relație cu mărimile statistice ale itemilor. În ceea ce privește viteza, testele pot fi clasificate în trei tipuri: teste de viteză pure, teste de putere pure, teste care combină viteza și puterea.

- *Testele pure de viteză* măsoară pur și simplu viteza cu care subiecții testați pot îndeplini o sarcină. Într-un test pur de viteză, dificultatea este manipulată în principal prin controlarea timpului. Aceste teste au itemi ale căror nivele

de dificultate sunt uniforme și corespund bine capacităților indivizilor care, probabil, vor fi testați, dar limitele de timp sunt atât de reduse, încât cei mai mulți subiecți nu pot rezolva complet toți itemii. Astfel, în cele mai multe cazuri, scorul total la un test pur de viteză este pur și simplu numărul de itemi rezolvați complet de persoana testată. Dacă persoanele testate termină toți itemii unui test pur de viteză, nu se determină capacitatea lor reală deoarece nu există nicio cale de a ști câți itemi ar mai fi putut rezolva complet dacă în test ar fi existat mai mulți itemi.

- *Testele pure de putere*, pe de altă parte, nu au limită de timp. În aceste teste dificultatea este manipulată prin creșterea sau descreșterea complexității itemilor. Intervalul lor de dificultate trebuie să fie suficient de mare pentru a cuprinde nivelele de capacitate pentru toți subiecții potențiali. În testele de putere, itemii sunt aranjați în ordine crescătoare a dificultății, astfel încât toți subiecții testați sunt capabili să rezolve cel puțin câțiva itemi. Un scor perfect la un test pur de putere sugerează că nivelul aptitudinii subiectului depășește nivelul de dificultate al celor mai dificili itemi. În astfel de cazuri, nivelul actual al aptitudinii subiectului este nedeterminat datorită plafonului insuficient al testului.
- Cele mai multe teste de aptitudine se încadrează între extremele continuumului viteză pură/putere pură. De obicei limitele lor de timp permit persoanelor testate să ajungă și să încerce să rezolve toți sau cei mai mulți dintre itemi. Așa cum am discutat anterior, intervalul specific și media nivelelor de dificultate ale itemilor testelor de capacitate depind de scopurile pentru care sunt utilizate testele.

În orice test temporizat cu atenție, valorile *p* și indicii capacității de discriminare a itemilor sunt funcții ale *poziției* acestora în test, și nu a dificultății intrinsece sau a validității discriminatorii. Se întâmplă așa deoarece la itemii din ultima parte a unui test în care viteza joacă un rol semnificativ ajung puțini dintre subiecții testați și cei care încearcă să rezolve acești itemi tind să fie ori cei mai capabili subiecți, ori cei care se grăbesc să termine testul, răspunzând la întâmplare. Prin urmare, este probabil ca indicii de dificultate și discriminare pentru itemii care apar la sfârșitul testelor de viteză să inducă în eroare și este necesară implementarea unor strategii speciale pentru a înțelege rolurile specifice ale vitezei și dificultății în astfel de teste.

Combinarea dificultății și discriminării itemilor

În lumina interrelației dintre dificultatea unui item și capacitatea de discriminare a unui item, construirea celor mai multe teste de aptitudine necesită analize care combină ambele caracteristici ale unui item. Există două abordări ale acestui scop. Metodele mai vechi constau în analizarea regresiei item-test, iar cele mai recente presupun utilizarea teoriei răspunsului la item (TRI).

Regresia item-test

Pentru a realiza analize ale regresiei item-test este necesar să calculăm proporția de indivizi de la fiecare nivel al scorului total care au răspuns corect la un anumit item. Figura 6.2 prezintă regresiile item-test pentru ambii itemi, reprezentate grafic utilizând datele din Tabelul 6.2. Graficele regresiei item-test combină informațiile despre dificultatea itemului și capacitatea de discriminare a unui item și ne permit să vizualizăm modul în care

Tabelul 6.2. Datele regresiei item–test pentru doi itemi

Scor total	Proporția de examinați care au răspuns corect la fiecare item	
	Item 1	Item 2
10	1,00	1,00
9	0,60	0,85
8	0,75	0,70
7	0,65	0,50
6	0,70	0,45
5	0,80	0,30
4	0,50	0,00
3	0,40	0,00
2	0,30	0,00
1	0,35	0,00

funcționează fiecare item în grupul testat. Dacă presupunem că mărimile statistice ale itemilor prezentate în Tabelul 6.2 s-au bazat pe eșantioane mari și reprezentative de subiecți, aceste date vor face posibilă evaluarea celor doi itemi și formularea anumitor concluzii despre aceștia, prezentate în paragrafele următoare.

- *Itemul 1 este mai ușor decât itemul 2 pentru că pragul său de 50% este mai scăzut.* Pragurile de 50% sunt reprezentate în Figura 6.2 prin linii punctate perpendiculare care au fost trasate în punctele în care graficele de regresie ale fiecărui item întâlnesc axa orizontală care prezintă scorul total la test. Pentru itemul 1, pragul de 50% se află în punctul în care scorul total este egal cu 4, în timp ce pentru itemul 2,

el se află în punctul în care scorul total este egal cu 7. Aceste date sugerează faptul că nivelul de aptitudine necesar pentru a avea șanse 50–50 de a răspunde corect la itemul 1 este mai scăzut decât nivelul de aptitudine necesar pentru a avea o șansă egală de a reuși la itemul 2.

- *Itemul 2 discriminează mai bine decât itemul 1.* Regresia item-test este mai abruptă pentru itemul 2 decât pentru itemul 1 și nu prezintă nicio inversare a direcției proporției de răspunsuri corecte la item pentru fiecare punct al scorului total.

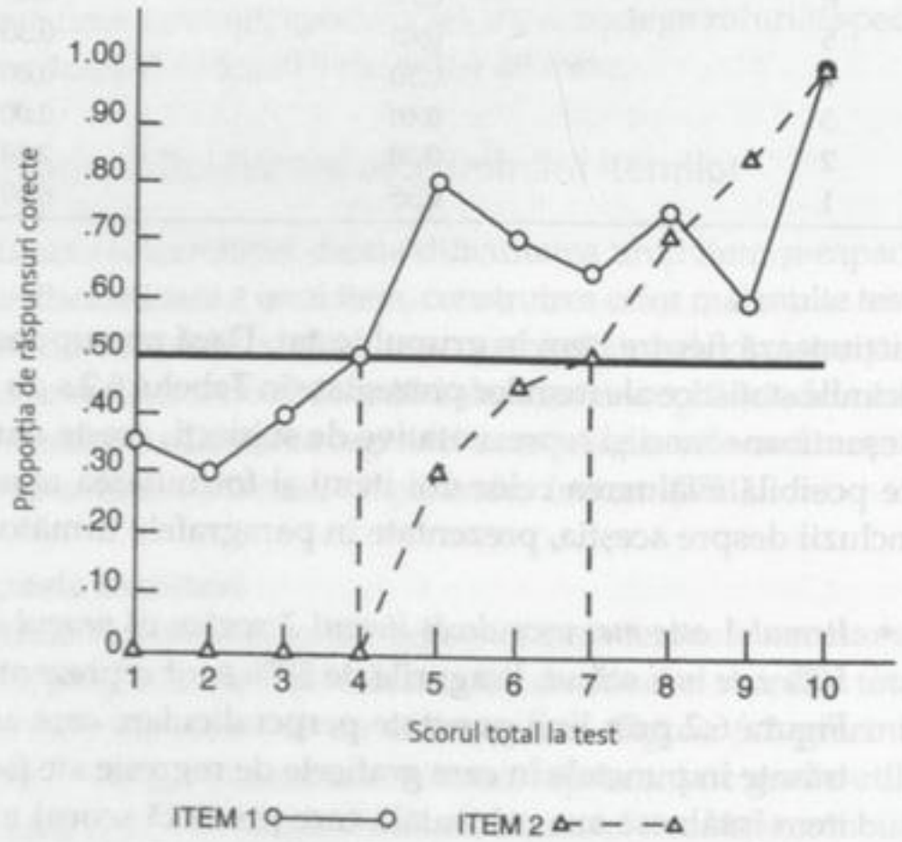

Figura 6.2 Regresia item-test pentru itemii 1 și 2 (vezi tabelul 6.2)

Prin opoziție, itemul 1 prezintă o regresie item-test treptată și patru inversări ale direcției sale (în punctele scorului total de 2,6,7 și 9). Din moment ce se presupune că scorul total la test reflectă nivelul aptitudinii subiectului testat, regresiile item-test din Figura 6.2 sugerează că relația dintre aptitudine și performanța la item este mai directă și mai stabilă pentru itemul 2 decât pentru itemul 1.

- *Este mai probabil ca la itemul 1 să se răspundă corect prin ghicirea răspunsului decât la itemul 2.* Această inferență se bazează pe faptul că proporția de răspunsuri corecte la Itemul 1 este destul de mare (0,35) chiar și pentru acei indivizi care au obținut un scor total de 1, cel mai scăzut scor la test. Dimpotrivă, niciun subiect cu scor sub 5 nu a fost capabil să răspundă corect (sau să ghicească răspunsul corect) la itemul 2.
- *Concluzie.* În general, examinarea datelor regresiei item-test prezentate în Tabelul 6.2 și Figura 6.2 sugerează faptul că (a) itemul 2 este mai dificil decât itemul 1, (b) itemul 2 pare a funcționa mai bine decât Itemul 1 în termenii puterii sale de a discrimina între indivizi cu scoruri scăzute și ridicate la un set ipotetic de 10 itemi de aptitudine și (c) itemul 2 este mai insensibil la ghicirea răspunsurilor corecte decât itemul 1.

Deși aceste analize ale regresiei item-test aduc informații, ele sunt destul de brute și destul de dependente de eșantioanele și seturile de itemi din care sunt obținute datele. Teoria răspunsului la item utilizează același tip de date empirice implicate în analizele regresiei item-test ca punct de plecare pentru forme de analize de item mult mai sofisticate și strategii de elaborare a unui test mult mai ambițioase.

Eticheta *teoria răspunsului la item (TRI)* se referă la o varietate largă de modele care pot fi utilizate pentru a construi și elabora teste noi și pentru a le evalua pe cele existente. Modelele TRI diferă prin formulele matematice pe care le utilizează, numărul de caracteristici ale itemului pe care le pot explica și numărul de trăsături sau dimensiuni ale aptitudinii pe care le specifică drept obiective ale măsurării. În plus, diferite metode sunt utilizate dacă datele itemului sunt dihotomice (corect-incorect, adevărat-fals, etc) sau politomice (i.e. constând din categorii multiple de răspuns). Procedurile din cadrul TRI sunt extensive și complexe. Până de curând, prezentările publicate ale acestor metode au fost prea dificile pentru a fi înțelese fără o cunoaștere solidă a matematicii și statisticii. Din fericire, în ultimii câțiva ani au fost publicate numeroase materiale excelente și mai accesibile despre tehnicile TRI. Info 6.4 enumeră o selecție a câtorva dintre cele mai utile resurse disponibile.

Teoria clasică a testării versus teoria răspunsului la item

Eticheta *teoria clasică a testelor (TCT)* este utilizată, în contradicție cu TRI, pentru a se referi la toate metodele psihometrice de construire și evaluare a testelor anterioare TRI. Metodele fundamentale ale TCT au fost elaborate în prima parte a secolului XX și au fost bine consolidate către mijlocul acestui secol. Probabil că au fost cel mai bine rezumate în volumul clasic al lui Gulliksen (1950) despre teoria testelor mentale, dar au fost descrise anterior în nenumărate alte surse. Principiile psihometrice și procedurile TCT au fost rafinate și extinse în mod continuu; ele

Info 6.4

Surse de informații despre Teoria răspunsului la item și testarea computerizată adaptativă

Cărți

- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Erlbaum.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.
- Wainer, H. (2000). *Computer adaptive testing: A primer* (ed. a 2-a). Mahwah, NJ:Erlbaum.

Resurse pe internet

Multe dintre resursele Teoriei răspunsului la item (TRI) disponibile anterior la Educational Resource Information Center (ERIC) Clearinghouse on Assessment and Evaluation pot fi găsite la http://edres.org/irt, o pagină web menținută de Lawrence Rudner, fost director al ERIC Clearinghouse on Assessment and Evaluation. Prin intermediul acestei pagini web se pot accesa numeroase materiale utile legate de TRI, cum ar fi:

- Un tutorial excelent despre teoria răspunsului la item, disponibil datorită Laboratorului de modelare TRI de la University of Illinois at Urbana-Champaign;
- A doua ediție a cărții clasice a lui Frank Baker *The Basics of Item Response Theory* (2001);
- Legături cu soft-uri TRI disponibile gratuit sau cu plată și colecții de articole și cărți despre TRI.

CAT Central este un site web cu diferite resurse pentru cercetare și aplicații ale testării adaptative computerizate (TAC), inclusiv informații fundamentale despre TAC, o bibliografie extinsă, o enumerare a programelor de testare majore care utilizează CAT și legături către celelalte resurse legate de CAT. CAT Central poate fi găsit la următoarea adresă:

- http://www.psych.umn.edu/psylabs/CATCentral

sunt încă larg utilizate și vor continua să fie utilizate în perioada care va urma. De fapt, cele mai multe cărți despre testarea psihologică — inclusiv cea de față — au de-a face în mare măsură cu TCT. Unele contraste majore între TCT și TRI au fost menționate pe scurt în Capitolul 3 în legătură cu problema echivalenței testelor. Cu toate acestea, intervalul și semnificația schimbărilor implicate în tranziția de la procedurile convenționale ale TCT și abordarea măsurătorii bazată pe modelul ce caracterizează TRI se extind mult dincolo de această temă.

În prezent, metodele TRI sunt utilizate într-un număr mult mai restrâns de instrumente decât metodele tradiționale ale TCT. Acest fapt se datorează pe de o parte premiselor pe care le cere TRI — cu privire la răspunsul la item, trăsături latente și relația lor — și pe de altă parte eforturilor mult mai extinse de colectare a datelor necesare pentru a calibra itemii utilizând modelele TRI. Mai mult, în contrast cu tehnicile bine stabilite, comparativ simple și larg utilizate ale TCT, metodele TRI sunt încă în dezvoltare, sunt mult mai sofisticate din punct de vedere matematic și nefamiliare chiar și pentru numeroși profesioniști din domeniul testării psihologice. Așa cum Embretson (1996, 1999) a exprimat extrem de clar, deși TCT și TRI au aceleași fundamente conceptuale, multe dintre regulile tradiționale de măsurare implicite în TCT trebuie revizuite sau abandonate atunci când modelele TRI sunt aplicate sarcinilor de măsurare.

Nu uitați!

Secțiunile despre teoria răspunsului la item (TRI) și testarea computerizată adaptativă (TAC) din Capitolul 3 oferă o introducere fundamentală în câteva dintre trăsăturile distinctive ale acestor abordări relativ noi ale măsurării psihologice. Pentru unii cititori poate fi util să se întoarcă la acele secțiuni anterioare ca preludiu pentru materialul despre acele teme prezentat în acest capitol.

Info 6.5

Teoria clasică a testării versus teoria răspunsului la item: perspective opuse cu privire la lungimea și fidelitatea unui test

Noile reguli de măsurare descrise de Embretson (1996, 1999) subliniază anumite diferențe cruciale între teoria clasică a testului (TCT) și teoria răspunsului la item (TRI). Printre acestea se află contrastul între regula veche conform căreia „testele mai lungi au fidelitate mai mare decât testele mai scurte" și noua regulă conform căreia „testele mai scurte pot avea fidelitate mai mare decât testele mai lungi"(p. 343). Cu alte cuvinte:

- Așa cum am discutat și în legătură cu fidelitatea determinată prin metoda înjumătățirii și formula Spearman-Brown (Capitolul 4), TCT susține că un număr mai mare de observații va produce rezultate mai fidele decât un număr mai redus de observații. Dacă lungimea unui test crește prin adăugarea de itemi paraleli, proporția variației reale din varianța erorii crește și ea și, ca urmare, crește și fidelitatea scorurilor testului. Astfel, pentru două teste comparabile de lungimi fixe (50 vs. 40 itemi), scorurile testului mai lung vor avea fidelitate mai mare decât scorurile obținute la testul mai scurt.
- În testarea adaptativă computerizată (TAC) pe care o permit metodele TRI, selecția itemilor este optimă pentru nivelele subiecților testați la trăsătura care este măsurată. Itemii neadecvați (cei care sunt prea ușori sau prea dificili pentru subiectul testat) sunt eliminați, ceea ce duce la un test mai scurt. Deoarece metodele TRI calibrează mai precis și informația obținută din fiecare răspuns la item, eroarea de măsură poate fi redusă și pot fi obținute scoruri fidele din mai puține răspunsuri la item, dar care conțin mai multe informații.
- Pentru explicații detaliate ale acestor noțiuni însoțite de exemple numerice și grafice vezi Embretson (1996, 1999) și Embretson și Reise (2000).

Info 6.5 prezintă câteva trăsături opuse ale celor două abordări.

Diferența fundamentală dintre TCT și TRI derivă din faptul că în TCT interesul se centrează în principal asupra scorului total al unui examinat la un test care reprezintă suma scorurilor la item, în timp ce în TRI — așa cum implică și numele — atenția se focalizează în principal asupra performanțelor examinatului la itemi individuali. În TRI preocuparea principală o constituie elaborarea atentă și calibrarea itemilor unui test în termenii informației pe care o oferă cu privire la un construct psihologic specific. Pentru a realiza această calibrare, TRI se bazează pe modele matematice ale relației dintre aptitudini — sau orice alte constructe neobservabile (i.e. trăsături latente) este destinat să măsoare un test — și răspunsurile la itemi individuali.

În general vorbind, scopurile TRI sunt (a) de a genera itemi care oferă un grad maxim posibil de informație cu privire la nivelele aptitudinii sau trăsăturii examinaților care răspund la itemi într-o manieră sau în alta, (b) de a oferi examinaților itemi care sunt ajustați la nivelele lor de aptitudine sau trăsătură și astfel (c) de a reduce numărul de itemi necesari pentru a localiza exact poziția fiecărui subiect la aptitudinea sau trăsătura latentă simultan cu minimizarea erorii de măsură. Reducerea numărului de itemi dintr-un test prin selectarea celor care sunt cei mai adecvați pentru nivelul de aptitudine al subiectului testat — fără o pierdere ulterioară a fidelității — este un obiectiv important al testării în grup. Aceast fapt este adevărat în special pentru programele de testare care sunt desfășurate pe scară largă, cum ar fi SAT. O reducere a numărului de itemi administrați duce la economie de timp, de bani și reduce la minim experiența de frustrare a subiecților testați atunci când sunt confruntați cu itemi care nu sunt potriviți nivelelor aptitudinii lor. Într-adevăr, în programele de testare pe scară largă, TAC care utilizează metode TRI înlocuiește

treptat testele de lungime fixă în format creion-hârtie sau administrate pe computer (Embretson & Reise, 2000).

Limite ale teoriei clasice a testării

Teoria clasică a testării și TRI diferă în multe alte privințe. Deși prezentarea completă a acestor diferențe depășește obiectivele acestui volum, câteva dintre acestea trebuie să fie menționate datorită semnificației lor cu privire la construirea unui test și analiza de itemi. O modalitate de a compara cele două metodologii este sublinierea limitelor TCT pe care TRI încearcă să le depășească. Deși unele dintre aceste puncte au fost menționate anterior în acest volum, ele sunt reluate aici mai amănunțit, cu referire specifică la comparația dintre TCT și TRI.

• Indicii TCT ai dificultății itemului și discriminării itemului sunt *dependenți de grupare*: valorile lor se pot schimba atunci când sunt calculați pentru eșantioane de subiecți care diferă de cele utilizate pentru analizele de item inițiale în privința anumitor aspecte ale constructului măsurat. Dimpotrivă, se presupune că estimările caracteristicilor unui item obținute prin metodele TRI sunt invariante și oferă o scală uniformă de măsurare care poate fi utilizată cu grupuri diferite.

• Pentru testele cu lungime fixă construite utilizând metode TCT, trăsătura sau estimările aptitudinii subiecților (i.e. scorurile) sunt *dependente de test*. Cu alte cuvinte, scorurile depind de itemii specifici selectați pentru a fi incluși în test. Prin urmare, compararea scorurilor obținute la diferite teste sau diferite seturi de itemi nu este posibilă cu excepția cazului în care sunt utilizate procedee de echivalare a testului, care deseori nu sunt fezabile (vezi Capitolul 3). Mai mult, chiar și atunci când sunt aplicate procedee

de echivalare, comparațiile care pot fi făcute se limitează la testele echivalate. În cazul TRI — dat fiind că datele de potrivesc modelului și că pot fi îndepinite anumite premise — estimările aptitudinilor sau trăsăturilor sunt independente de setul particular de itemi administrat examinaților. Estimările trăsăturii sunt legate de probabilitățile tiparelor de răspuns la item ale celor examinați.

• În metodologia TCT, fidelitatea scorurilor (i.e. estimărea trăsăturii sau aptitudinii) este stabilită prin intermediul erorii standard de măsură (*SEM*) care se presupune a fi la fel de mare pentru toți examinații (vezi Capitolul 4). De fapt, din moment ce fidelitatea scorurilor depinde de cât de potriviți sunt itemii testului pentru nivelele trăsăturii sau aptitudinii examinatului, și din moment ce nivelele trăsăturii nu sunt egale pentru diferiți examinați, această premisă nu este adevărată în cazul testelor tradiționale. Pe de altă parte, atunci când metodologia TRI este combinată cu proceduri de testare adaptativă, erorile standard ale estimărilor trăsăturii sau aptitudinii rezultate din administrarea unui test depind de setul particular de itemi selectat pentru fiecare examinat (vezi Info 6.5). Drept consecință, aceste estimări variază adecvat la diferite nivele ale dimensiunilor trăsăturii și transmit informații mai precise cu privire la fidelitatea măsurării.

Caracteristicile esențiale ale teoriei răspunsului la item

Deoarece cele mai multe modele TRI utilizate în prezent sunt modele unidimensionale, prezentarea de față se va limita la acestea. *Modelele TRI unidimensionale* presupun că (a) itemii care alcătuiesc un test sau un segment de test măsoară o singură trăsătură și (b) răspunsurile la item date de subiecți depind numai de poziția acestora la trăsătura măsurată. Așa cum sugerează Info 6.6, dintr-o

perspectivă realistă, niciuna dintre aceste premise ale modelelor unidimensionale nu poate fi îndeplinită în totalitate. Totuși, atunci când toți itemii dintr-un test sau dintr-un segment de test sunt destinați să măsoare o singură trăsătură predominantă, premisele modelelor unidimensionale pot fi îndeplinite suficient de adecvat pentru ca modelul să funcționeze.

Info 6.6

De ce eșantioanele de comportament obținute prin intermediul itemilor unui test sunt atât de complexe?

- *Indiferent de constructul pe care itemii testului intenționează să-l măsoare, aceștia presupun întotdeauna dimensiuni multiple.* Mai întâi, pentru a răspunde la *orice* sarcină a testului, este necesară o anumită aptitudine de a acorda atenție stimulilor testului, cum ar fi un minimum de memorie de scurtă durată. Mai mult, toți itemii testului presupun un conținut, format și mediu specifice și necesită un set specific de abilități cognitive. De exemplu, în funcție de modul său de prezentare și răspuns, un simplu item de vocabular poate presupune citire, scriere, înțelegere orală, exprimare verbală, capacitate de raționament logic și cunoștințe de etimologie, fără a mai menționa atenția, memoria și, posibil, viteza de răspuns.
- *Subiecții testați sunt ființe complexe și unice.* Ei aduc o combinație de factori — cum ar fi zestrea genetică, istoriile de viață, aptitudini dezvoltate, trăsături, obiceiuri și stări fiziologice și emoționale tranzitorii — relevante pentru sarcinile testului. Deoarece răspunsurile la itemii testului depind de amestecul unic al tuturor elementelor pe care subiectul testat le aduce în sarcină, astfel de răspunsuri nu sunt niciodată echivalente sub toate aspectele. De exemplu, itemii testului care cer o serie de calcule aritmetice prezintă o problemă mai mare pentru un subiect care se teme de matematică decât pentru unul care nu se teme, chiar dacă amândoi sunt la fel de capabili să realizeze calcule într-o situație non-testare.

În următoarele câteva paragrafe sunt sintetizate unele trăsături comune celor mai multe modele TRI utilizate în prezent pentru a le oferi cititorilor o idee generală a modului în care metodologia TRI este aplicată calibrării datelor unui item al testului. Pentru concizie și simplitate, această prezentare evită utilizarea formulelor și conceptelor matematice care nu sunt esențiale pentru o înțelegere elementară a ideilor fundamentale ale TRI. Cititorii interesați pot găsi prezentări mai extinse ale acestor metode și explicații ale bazelor lor matematice în sursele enumerate în Info 6.4.

• În TRI, modelele se bazează pe premisa că performanța unei persoane la orice item al testului depinde de una sau mai multe trăsături sau aptitudini și poate fi prezisă de acestea. Modelele caută să specifice relația așteptată între răspunsurile examinaților la itemi (observabile) și trăsăturile (neobservabile) care determină răspunsurile. Deoarece presupun predicții, modelele TRI pot fi evaluate (i.e. confirmate sau respinse) în funcție de cât de bine se potrivesc ele datelor derivate din răspunsurile la itemii testului.

• Metodele TRI utilizează datele răspunsurilor la test și la item din eșantioane mari despre care se știe că diferă în privința aptitudinii sau trăsăturii de personalitate pe care testul aflat în construcție este destinat să o evalueze. Astfel de eșantioane nu trebuie să fie reprezentative pentru o populație definită, ci trebuie să includă grupuri de indivizi aflați la nivele diferite pe continuumul trăsăturii sau aptitudinii. În plus, itemii administrați inițial trebuie construiți sau selectați cu atenție față de potențialul lor de indicatori ai trăsăturii ce urmează a fi evaluată.

• După ce scorurile la item și la test sunt colectate, acestea sunt utilizate pentru a obține estimări ale parametrilor itemilor ce vor plasa examinații și itemii pe o scală comună pentru dimensiunea aptitudinii sau trăsăturii. *Parametrii itemului* sunt valori numerice care specifică forma relației dintre aptitudinile sau trăsăturile măsurate și probabilitatea unui anumit răspuns la item. De exemplu, parametrii *dificultății itemului* exprimă dificultatea unui item în termenii poziției subiectului pe scala aptitudinii pentru care probabilitatea de a rezolva corect itemul este de 50%. Tabelul 6.3 prezintă un set redus, ipotetic, de date brute obținute cu 10 itemi scorați dihotomic, administrat unui număr de 10 indivizi (de la A la J). Deși un exemplu realist ar include un eșantion de subiecți mult mai mare și mai divers — probabil grupați în două categorii pe baza scorurilor lor totale, și nu individual —, datele din Tabelul 6.3 ilustrează tipul de informație care ar putea să fie utilizată pentru a estima parametrii dificultății itemului în raport cu nivelele aptitudinii. Parametrii itemului sunt obținuți prin diferite procedee care necesită utilizarea programelor de computer specializate (vezi Embretson & Reise, 2000, cap. 13). Aceste procedee presupun funcții matematice *neliniare*, cum ar fi funcțiile logistice care determină curbe caracteristice ale itemului (vezi mai jos). Modelele matematice neliniare sunt necesare deoarece modelul regresiei liniare, prezentat în Capitolele 2 și 5, nu este potrivit pentru a descrie modul în care schimbările nivelelor trăsăturilor este legat de schimbările probabilității de a răspunde la un item într-un mod specific.

• *Curba caracteristică a itemului (CCI)* este reprezentarea grafică a unei funcții matematice care pune în relație probabilitățile de răspuns la item cu nivelele trăsăturii, dați fiind parametrii

Tabelul 6.3 Exemplu ipotetic de date brute pentru itemi și persoane utilizate în estimarea parametrului TRI

	Item										
Persoană	1	2	3	4	5	6	7	8	9	10	Total
A	1	1	1	1	0	1	1	1	1	1	8
B	0	0	1	1	1	1	0	0	0	0	4
C	0	0	1	1	1	1	0	0	0	0	4
D	1	1	0	0	1	0	0	0	1	1	5
E	1	1	1	1	1	1	1	1	1	1	10
F	1	1	1	1	0	0	0	0	1	1	6
G	1	1	0	1	0	1	0	1	1	1	7
H	0	1	1	1	0	0	0	0	0	0	3
I	0	0	1	0	0	0	0	1	1	1	4
J	0	1	1	1	0	0	0	0	1	1	5
Total	5	7	8	8	4	4	2	4	7	7	

Notă: 1 și 0 din fiecare celulă a tabelului arată dacă persoanele de la A la J au răspuns corect sau incorect la fiecare dintre cei 10 itemi. Scorurile totale ale subiecților, din ultima coloană, pot fi utilizate pentru a calcula estimările aptitudinii; scorurile totale ale itemului, de pe ultimul rând, pot fi utilizate pentru a calcula dificultățile itemilor.

specificați ai itemului. De exemplu, CCI a unui item dintr-un test de aptitudine dihotomic exprimă vizual relația așteptată dintre nivelul de aptitudine și probabilitatea de a răspunde corect la item. În cazul itemilor unui test de personalitate, CCI prezintă grafic relația așteptată dintre nivelele trăsăturii de personalitate și probabilitatea de a răspunde la un item într-o manieră specifică. CCI ipotetice prezentate în Figura 6.3 exemplifică trei dintre cele mai comune modele logistice

unidimensionale pentru date dihotomice ale răspunsului la item. Graficul A din figura 6.3 prezintă CCI pentru un model logistic cu parametru unic, cunoscut drept *modelul Rasch* în onoarea matematicianului care l-a elaborat (Rasch, 1960/1980). Tablourile B și C ale figurii 6.3 ilustrează modelele logistice cu doi și, respectiv, trei parametri.

- *Graficul A din Figura 6.3* prezintă CCI pentru doi itemi care diferă numai în privința dificultății. Itemul 1 este mai ușor decât Itemul 2. Poziția parametrului dificultate (i.e. nivelul de aptitudine asociat cu o probabilitate de reușită de 50% sau 0,50) este mai scăzută pentru itemul 1 (X_1) decât pentru itemul 2 (X_2). Din moment ce pantele celor două curbe sunt aceleași, putem deduce că cei doi itemi funcționează la fel de bine în termenii relației dintre aptitudine și probabilitatea de reușită la scala de aptitudine.
- *Graficul B din Figura 6.3* prezintă CCI pentru doi itemi care diferă în privința a doi parametri, dificultate și discriminare. În această situație, nivelul aptitudinii asociat cu o probabilitate de succes de 50% este ceva mai înalt pentru itemul 1 (X_1) decât pentru itemul 2 (X_2). Mai mult, pantele celor două curbe — care prezintă raportul dintre modificarea abilității și modificarea probabilității de succes pentru fiecare item — sunt diferite și se intersectează la un anumit punct. Această configurație sugerează că cei doi itemi funcționează diferit sub aspectul relației lor cu aptitudinea și nu discriminează la fel de bine în toate punctele scalei de aptitudine. Într-un test cu adevărat unidimensional, itemii cu CCI care se intersectează — cum sunt cei din Graficul B — nu sunt dezirabili.

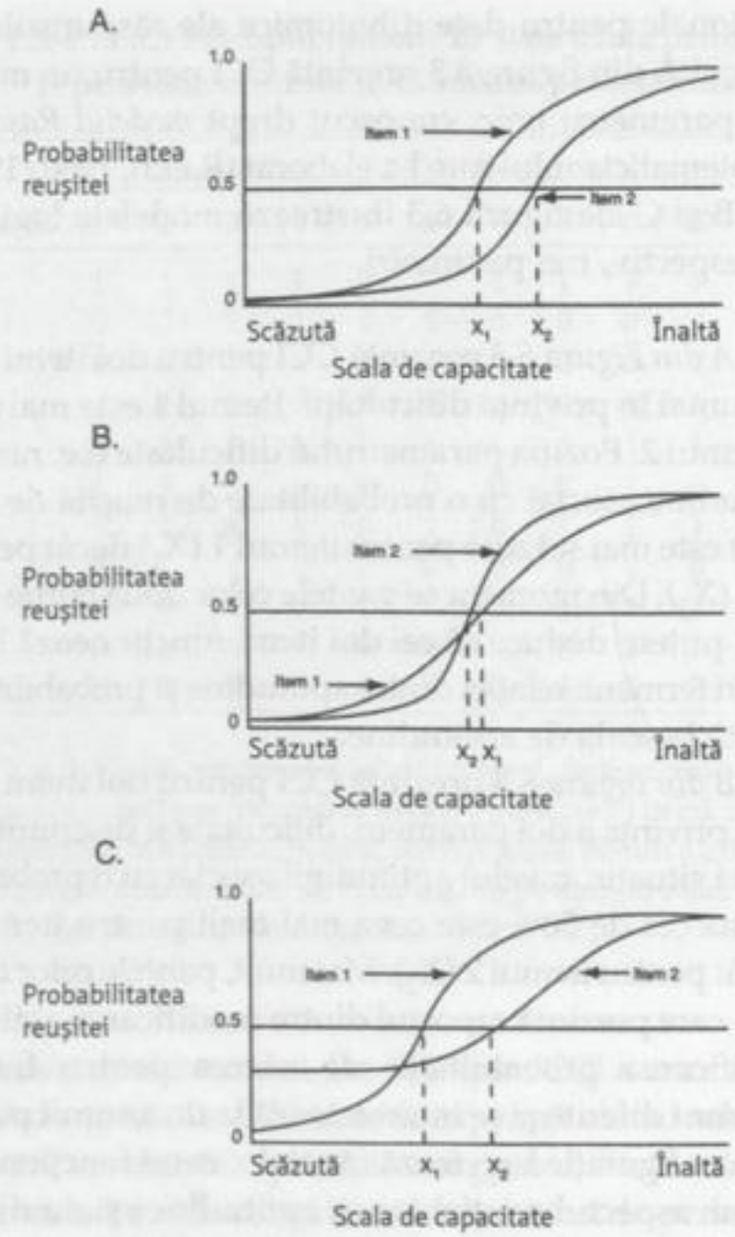

Figura 6.3 Curbele caracteristice ale itemilor: A, model cu parametru unic; B, model cu doi parametri; și C, model cu trei parametri

- *Graficul C din Figura 6.3* prezintă CCI pentru doi itemi care diferă din perspectiva a trei parametri și anume dificultatea, discriminarea și probabilitatea de reușită din întâmplare (sau ghicire). CCI pentru itemul 1 are o pantă mai mică decât curba pentru Itemul 2 și prezintă o creștere constantă a probabilității de succes pe măsură ce nivelele aptitudinii cresc până la un anumit punct. Dimpotrivă, itemul 2 în mod clar nu discriminează la fel de bine ca itemul 1 între indivizi aflați la diferite nivele de aptitudine: CCI pentru itemul 1 prezintă o relație mai puțin pronunțată între nivelul aptitudinii și probabilitatea reușitei. Observăm, de asemenea, că CCI pentru itemul 2 prezintă o probabilitate de succes destul de mare chiar și la capătul inferior al spectrului de valori ale aptitudinii. Acest fapt sugerează că subiecții testați aflați la nivelele scăzute ale aptitudinii sunt capabili să ghicească răspunsul corect la itemul 2. Mai mult, probabilitatea de succes de 50% este asociată cu un nivel mai ridicat al aptitudinii (X_2) pentru itemul 2. În mod clar, un item cu CCI asemănător cu CCI pentru itemul 2 din Graficul C va fi mai puțin eficient decât itemul 1 din perspectiva măsurării.
- Așa cum se întâmplă cu orice model teoretic, măsura în care premizele modelelor TRI sunt îndeplinite poate fi determinată

Nu uitați!

În teoria răspunsului la item (TRI), se presupune că parametrii unui item sunt *invarianți* pentru populație, însemnând că ei trebuie să fie stabili chiar dacă sunt calculați pentru grupuri care diferă în privința aptitudinii sau trăsăturii măsurate. Astfel, spre deosebire de statisticile analizei de item descrise anterior în acest capitol, parametrii TRI nu sunt legați de performanța unui grup de referință dat.

comparând predicțiile lor cu datele empirice și evaluând mărimea și semnificația oricăror discrepanțe existente între datele observate și predicțiile făcute pe baza acestor modele. Dacă potrivirea dintre așteptările bazate pe modelul CCI și performanța examinaților la un item al testului este suficient de mare, parametrii TRI sunt utilizați pentru a obține funcția de informare a itemului.

- *Funcția de informare a unui item* reflectă contribuția pe care un item o aduce la estimarea unei trăsături sau aptitudini pe un continuum al trăsăturii sau aptitudinii. Funcția de informare a itemului ajută la a decide dacă și unde să fie încorporați itemii într-un test, din perspectiva obiectivelor testului. Estimările nivelului trăsăturii care localizează subiecții pe dimensiunea trăsăturii sunt obținute din tiparele lor specifice de reușită sau eșec la o serie de itemi. În TRI, *funcția de informare a testului* care reprezintă suma funcțiilor de informare ale itemilor corespunde noțiunii de fidelitate a scorurilor din TCT (vezi Capitolul 4). Funcțiile de informare a testului sunt calculate și utilizate pentru a obține erori standard ale estimării pentru fiecare nivel al scalei trăsăturii sau aptitudinii. La rândul lor, aceste erori standard creează benzi de încredere pentru estimările capacității într-o manieră similară modului în care erorile standard de măsură din TCT sunt utilizate pentru a crea benzi de încredere pentru scorurile obținute.

Așa cum se poate determina din comentarea CCI prezentate în Figura 6.3, chiar și modelele multidimensionale TRI exemplificate în această figură pot deveni destul de complexe pe măsură ce crește numărul de parametri cuprinși în model. Aplicarea modelelor TRI la elaborarea instrumentelor destinate evaluării

unor constructe intelectuale și de personalitate mai largi și mai controversate este un proiect mult mai dificil (vezi Reise & Henson, 2003). Modelele multidimensionale TRI — care presupun că două sau mai multe trăsături contribuie la răspunsurile la item — sunt utilizate în prezent pentru a explora și explica constructe mai complexe și multifațetate. Unele dintre aceste modele mai noi și mai complicate sunt descrise de Embretson și Reise (2000).

Corectitudinea itemului

În general vorbind, există multe moduri prin care itemii unui test, ca și testele, pot prezenta aspecte preferențiale sau pot fi incorecți față de unul dintre subiecții testați sau față de grupuri de subiecți. Atât timp cât este vorba despre teste, posibilitatea existenței unor aspecte preferențiale poate fi investigată stabilind dacă scorurile la test au aceeași semnificație pentru membrii diferitelor subgrupuri din populație (vezi Capitolul 5). Pe de altă parte, problema *corectitudinii* unui test este un aspect mult mai complex și mai controversat. Deși există un acord general cu privire la faptul că utilizarea incorectă a testelor trebuie evitată, ceea ce reprezintă exact corectitudinea în testarea psihologică este o chestiune supusă unei dezbateri considerabile (AERA, APA, NCME, 1999, pp. 74–76). Cu toate acestea, utilizatorii testelor au o responsabilitate majoră în implementarea practicilor corecte de testare prin luarea în considerare a adecvării instrumentelor la scopurile propuse și la potențialii subiecți (vezi Capitolul 7).

La nivelul itemilor testului, întrebările cu privire la aspectele preferențiale și incorectitudine sunt mai circumscrise și, de obicei, sunt abordate când testul este încă în faza de construcție. În acest scop, itemii testului sunt analizați calitativ și cantitativ de-a lungul

procesului de construire a testului. Bineînțeles, măsura în care itemii unui test trec prin aceste revizuiri este legată de scopul intenționat al unui test. Se acordă atenție specială eliminării oricărui element preferențial sau incorect posibil sau din itemii testelor de aptitudine care urmează să fie utilizate în luarea unor decizii cu consecințe semnificative pentru subiecții testați.

Analiza calitativă a aspectelor preferențiale ale unui item

Evaluarea calitativă a itemilor testului din punctul de vedere al corectitudinii se bazează pe proceduri de evaluare realizate prin intermediul unor mese rotunde la care participă indivizi eterogeni din punct de vedere demografic, selectați datorită sensibilității lor la astfel de aspecte și, de preferat, datorită expertizei lor și în ariile acoperite de test. În mod tipic, aceste revizuiri au loc în două etape. În timpul etapei inițiale a construirii testului, când itemii sunt scriși sau generați, aceștia sunt examinați pentru (a) a selecta inițial orice descrieri stereotipe ale unor subgrupuri identificabile din populație, (b) a elimina itemi al căror conținut poate fi o ofensă la adresa membrilor grupurilor minoritare și (c) a se asigura faptul că diverse subgrupuri sunt reprezentate adecvat în materialele conținute într-un grup de itemi. În această revizuire inițială, indivizii familiari cu obiceiurile lingvinstice și culturale ale subgrupurilor specifice ce s-ar putea afla printre potențialii subiecți testați ar trebui, de asemenea, să identifice conținutul itemilor care ar putea funcționa în beneficiul sau în detrimentul unui grup specific, astfel încât acel item să poată fi revizuit. Al doilea stadiu al revizuirii calitative a unui item are loc mai târziu în procesul construirii testului, după ce itemii au fost administrați și datele performanței la item au fost analizate separat pentru diferite subgrupuri. În acest stadiu, sunt examinați

itemii care, pentru un anumit subgrup, prezintă diferențe în privința indicilor de dificultate, de discriminare sau a ambilor indici pentru a identifica motivele acestor diferențe, iar acești itemi sunt fie revizuiți, fie înlăturați.

Analiza cantitativă a aspectelor preferențiale ale itemilor

Evaluarea cantitativă a aspectelor preferențiale ale itemilor a fost legată uneori pur și simplu de diferențele dificultății relative a itemilor testului pentru indivizi din diverse grupuri demografice. Cu toate acestea, interpretarea aspectelor preferențiale ale itemilor este considerată drept naivă de diverși profesioniști din domeniul testării psihologice care nu consideră că diferențele dificultății relative ale unui item pentru diferite grupuri sunt o dovadă suficientă că itemul prezintă aspecte preferențiale (vezi Drasgow, 1987). Din perspectivă psihometrică, se consideră că un item este preferențial numai dacă probabilitatea de a răspunde la item într-o manieră specifică diferă în cazul indivizilor cu aceeași poziție la o anumită trăsătură. În testele de aptitudine, de exemplu, se poate considera că există aspecte preferențiale atunci când persoanele care au același nivel la o aptitudine, dar aparțin unor grupuri demografice diferite, au probabilități diferite de a reuși la un item. Astfel, în literatura despre testarea psihologică, orientarea preferențială a unui item este descrisă mai adecvat drept *funcționare diferențială a itemului (FDI)*, o etichetă care denotă mai pregnant situațiile în care relația dintre performanța la item și constructul evaluat de un test diferă în două sau mai multe grupuri.

Procedurile clasice pentru analiza cantitativă a FDI presupun comparații ale statisticilor dificultății itemului și discriminării itemului pentru grupuri diferite. De exemplu, dacă un item al

unui test corelează slab cu scorul total obținut la test (i.e. discriminare redusă) și este mai dificil pentru femei decât pentru bărbați, în mod evident itemul ar fi suspect și ar trebui să fie înlăturat. Cu toate acestea, analiza FDI prin intermediul comparațiilor simple ale corelațiilor item-test și a valorilor *p* pentru diferite grupuri este complicată de faptul că grupuri de diferite tipuri (grupuri de gen, etnice, socioeconomice etc.) diferă deseori sub aspectul performanței și variabilității lor medii, mai ales la testele de aptitudine. Atunci când diferențele de grup de acest tip apar în distribuțiile scorurilor la test, (a) statisticile dificultății itemului se confundă cu diferențele valide dintre grupuri ale aptitudinii pe care testul o măsoară și (b) indicii de corelație ai discriminării itemului sunt influențați de diferențele variabilității din grupurile comparate. Din cauza acestori factori care complică lucrurile, statisticile analizei de item tradiționale nu s-au arătat foarte utile în detectarea funcționării diferențiale a itemului.

Funcționarea diferențială a itemului

Evaluarea adecvată și studierea FDI recesită metode specializate și au fost propuse numeroase metode. Una dintre cele mai utilizate în mod obișnuit este tehnica Mantel-Haenszel (Holland & Thayer, 1988) care extinde procedeele tradiționale de analiză de item. În acest tip de analiză, fiecare dintre grupurile în chestiune (e.g. grupuri majoritare și minoritare) sunt divizate în subgrupuri pe baza scorului total la test, iar performanța la item este evaluată în subgrupuri comparabile. Deși această metodă este mai rafinată decât comparația simplă a statisticilor analizei de item pentru grupuri, procedura Matel-Haenszel se bazează încă pe un criteriu intern (scor total) care poate fi insensibil la diferențele de

funcționare a itemului în grupuri, iar capacitatea sa de a detecta FDI este dependentă fundamental de utilizarea unor grupuri foarte mari (Mazor, Clauser, & Hambleton, 1992).

Teoria răspunsului la item oferă un fundament mult mai bun pentru investigarea FDI decât metodele teoriei clasice a testării. Pentru a stabili dacă indivizii din diferite grupuri cu nivele similare ale trăsăturii latente au rezultate diferite la un item este necesar să localizăm persoanele din două sau mai multe grupuri pe o scală comună de aptitudine. Procedurile TRI pentru atingerea acestui scop încep prin identificarea unui set de itemi ancoră care nu prezintă nicio FDI în grupurile de interes. Când se realizează aceasta, itemi suplimentari pot fi evaluați pentru FDI prin compararea estimărilor parametrilor itemului și a CCI obținute separat pentru fiecare grup. Dacă parametrii și CCI pentru un item dat, determinate pentru două grupuri, sunt identice, s-ar putea deduce că itemul funcționează la fel de bine pentru ambele grupuri. Nu este surprinzător faptul că procedeele TRI au devenit metode preferate pentru detectarea FDI (vezi Embretson & Reise, 2000, cap. 10, pentru detalii suplimentare).

Aplicații ale teoriei răspunsului la item

Așa cum am indicat în Capitolul 3, utilizarea metodelor TRI în dezvoltarea unui test și în calibrarea itemului nu împiedică interpretarea scorurilor unui test în raport cu norma sau cu criteriul. De fapt, datorită metodelor sale mai rafinate de calibrare a itemilor unui test și de determinare a erorii de măsură, TRI poate extinde interpretarea scorurilor obținute la test. Deși TRI nu poate oferi soluții pentru toate problemele măsurării psihologice, deja a ajutat la abordarea mult mai obiectivă și mai disciplinată a

elaborării testelor în ariile în care a fost aplicată. În prezent, metodele TRI au fost aplicate cel mai mult în elaborarea testelor adaptative computerizate utilizate în programe de testare pe scară largă, cum ar fi SAT și ASVAB. Elaborarea testelor de acest tip necesită contribuția unor indivizi care, pe lângă cunoștințe ale ariei de conținut acoperită de teste, au o expertiză tehnică considerabilă în matematică și programare. Aplicații mai limitate ale metodelor TRI au fost utilizate pentru o perioadă de timp. De exemplu, evaluarea parametrilor dificultății itemului prin metodele TRI a devenit destul de obișnuită în dezvoltarea bateriilor de aptitudini și achiziții de cunoștințe, cum ar fi Differential Ability Scales, scalele Wechsler, testele Wide Range Achievement și testele Woodcock. Modelele teoriei răspunsului la item sunt utilizate, de asemenea, din ce în ce mai mult în evaluarea FDI în testele cognitive. Deși metodele TRI sunt promițătoare și pentru domeniul testării personalității, aplicarea lor în această arie a fost mai limitată decât în testarea aptitudinilor (Embretson & Reise, 2000, cap. 12).

Explorări în elaborarea și scorarea itemilor

Revoluția din domeniul tehnologiei computerizate, alături de ritmul rapid al progreselor din teoria și metodologia din psihologie permit o explorare aproape nelimitată a tehnicilor inovatoare care pot fi aplicate problemelor legate de măsurare. În finalul acestui capitol sunt prezentate două aplicații promițătoare ale tehnologiei computerizate la domeniul itemilor testului.

Progrese recente în generarea itemilor

Descompunerea sarcinii și metodele de analiză a protocolului testării introduse de psihologia cognitivă au condus la progrese semnificative în explorarea și clarificarea proceselor, strategiilor și volumului de cunoștințe implicate în performanța la itemii testului (vezi Capitolul 5). De fapt, încă din anii '80, aceste progrese — alături de pașii concomitenți în TRI și tehnologie — au fost aplicate în generarea computerizată a itemilor testului pentru construirea testelor de aptitudine și de achiziții de cunoștințe. Această metodologie se află încă la început, comparativ vorbind, deoarece specificațiile necesare pentru a crea reguli prin care programele de calculator pot genera itemi ai testului trebuie să fie considerabil mai detaliate decât pentru metodele tradiționale de generare a itemilor și necesită un nivel mai înalt al fundamentului teoretic. Cu toate acestea, generarea computerizată a itemului a fost deja implementată în elaborarea unor instrumente pentru arii — cum ar fi evaluarea cunoștințelor de matematică și testarea aptitudinilor în aviație — în care (a) există modele cognitive ale performanței, (b) constructele ce urmează a fi examinate pot fi prezentate în termenii unei sintaxe logice și (c) dificultatea itemilor poate fi determinată prin referenți obiectivi. Fără îndoială, tehnicile de generare a itemilor vor continua să fie urmărite activ de cercetători datorită numeroaselor avantaje pe care le prezintă în termenii eficienței și economiei, precum și pentru potențialul lor în aplicațiile pedagogice (Irvine & Kyllonen, 2002).

Scorarea automată a eseurilor

O inovație semnificativă în efortul de a standardiza scorarea eseurilor este elaborare unei tehnologii computerizate pentru *scorarea automată a eseurilor (SAE)*. Evaluarea prozei scrise prin software s-a desfășurat în ultimele câteva decenii și, asemenea generării computerizate a itemilor, a fost facilitată de progresele din psihologia cognitivă și din științele computaționale. Deși încă

în primele sale stadii, SAE prezintă mari promisiuni nu numai ca mijloc de a spori fidelitatea și validitatea scorurilor, dar și ca instrument de instruire (Shermis & Burstein, 2003).

Testați-vă singuri

1. **Procedurile implicate în analiza de item sunt realizate în primul rând de**
   a) cei care construiesc testul.
   b) utilizatorii testului.
   c) subiecții testați.
   d) cei care administrează testul.

2. **În mod tipic, analiza calitativă a itemului are loc**
   a) după ce un test este standardizat.
   b) simultan cu validarea încrucișată.
   c) după ce este generat grupul de itemi.
   d) chiar înainte de publicarea testului.

3. **Itemii cu alegere forțată aparțin categoriei**
   a) itemilor cu răspuns construit.
   b) itemilor cu răspuns la alegere.
   c) nici (a), nici (b).

4. **Care dintre următoarele avantaje *nu* este întâlnit în cazul itemilor cu răspuns la alegere în raport cu itemii cu răspuns construit? Itemii cu răspuns la alegere**
   a) sunt mai puțin predispuși la erori de scorare.
   b) utilizează mai eficient timpul de testare.
   c) sunt mai ușor de cuantificat.
   d) sunt mai ușor de pregătit.

5. **Indicii de discriminare sunt mărimi statistice utilizate în principal pentru a evalua**
   a) validitatea unui item.
   b) corectitudinea itemilor.
   c) fidelitatea itemului.
   d) dificultatea.

6. **Dacă cineva dorește să producă un test care ar duce la o diferențiere maximă între subiecții testați, ar urmări un nivel mediu al dificultății (valoare *p*) de**
   (a)1,00.
   (b) 0,75.
   (c) 0,50.
   (d) 0.

7. **Pearson *r* este coeficientul de corelație cel mai des utilizat pentru a calcula indicii relației dintre performanța la item și poziția la criteriu.** Adevărat sau fals?

8. **Pentru un test pur de viteză, indicii de dificultate și discriminare a itemului utilizați de obicei sunt _____ pentru testele pure de putere.**
   (a) mai puțin adecvați decât
   (b) mai adecvați decât
   (c) la fel de potriviți ca

9. **Care dintre următoarele afirmații *nu* este unul dintre obiectivele fundamentale pentru care se utilizează teoria răspunsului la item (TRI)?**
   a) A oferi maximul de informație despre nivelele trăsăturii examinaților
   b) A oferi examinaților itemi potriviți pentru nivelul lor
   c) A crește numărul de itemi incluși într-un test
   d) A minimiza eroarea de măsurare

10. **Până acum teoria răspunsului la item a fost aplicată mai puțin în aria testării _____.**
    a) achizițiilor de cunoștințe
    b) aptitudinilor
    c) cognitivă
    d) personalității

Răspunsuri corecte: 1. a; 2. c; 3. b; 4. d; 5. a; 6. c; 7. Fals; 8. a; 9. c; 10. d.

Aspecte esențiale ale utilizării testelor

Înainte de a fi utilizate, testele pot fi evaluate numai din punct de vedere științific și tehnic de către persoane cu expertiza necesară în construirea testelor, principiile psihometrice și aspectele comportamentului pe care testul încearcă să le evalueze. După realizarea testării — prin procesele de selecție, administrare și scorare — rezultatele testelor trebuie să fie evaluate, interpretate și comunicate într-o manieră adecvată scopului în care acestea urmează să fie utilizate de profesioniști care au cunoștințe despre contextul în care are loc testarea și despre aspectele tehnice și psihologice presupuse de o situație dată. În sens restrâns, evaluarea aplicării testelor psihologice și educaționale presupune considerații referitoare la abilitatea cu care utilizatorul testului folosește aceste instrumente și la adecvarea lor la testarea anumitor persoane. În sens larg, în măsura în care testarea dobândește o semnificație practică în viața indivizilor, utilizarea unui test trebuie să fie evaluată în lumina valorilor societății și a priorităților politice. În acest context mai larg, testarea poate deveni extrem de controversată și utilizatorii testelor trebuie să se bazeze pe codurile de etică personală și profesională pentru a stabili dacă sunt dispuși să utilizeze testele într-un anumit mod.

Practicile adecvate de testare sunt reglate de principii etice și standarde promulgate de fiecare dintre profesiile în care se utilizează teste psihologice și educaționale (American Counseling Association, 1995; AERA, APA, NCME, 1999; APA, 2002; National Association of School Psychologists, 2000). În ultimele decenii, neliniștile din ce în ce mai mari legate de posibilitatea utilizării greșite a testelor au făcut ca profesiile și organizațiile implicate în testare să se angajeze în pregătirea și difuzarea informațiilor despre calificările utilizatorilor de teste. Unul dintre eforturile majore în această direcție a fost promovat de APA și a dus la publicarea unui set explicit de linii directoare pentru a informa toate părțile implicate cu privire la calificările pe care APA le consideră importante pentru utilizarea competentă și responsabilă a testelor (Turner, DeMers, Fox, & Reed, 2001). Info 7.1 oferă un scurt rezumat al acestora. În plus, *Testing Standards* (AERA, APA, NCME, 1999) include un capitol care subliniază responsabilitățile utilizatorilor de teste psihologice, precum și capitole separate dedicate aspectelor legate de (a) corectitudinea în testare și utilizarea testelor, (b) testarea indivizilor din diverse medii lingvistice, (c) testarea indivizilor cu dizabilități, (d) testarea psihologică și evaluarea, (e) testarea și evaluarea educațională, (f) testarea în angajare și certificare și (g) testarea în evaluarea unui program și în domeniul politicilor publice.

De asemenea, documentele care oferă indicații suplimentare pentru utilizarea testelor ca parte a luării deciziilor deosebit de importante pentru studenți (U.S. Department of Education, Office for Civil Rights, 2000) și principiile pentru validarea și utilizarea procedurilor de selecție a personalului (Society for Industrial and Organizational Psychology [SIOP], 2003) sunt disponibile și merită consultate înainte de a iniția o activitate în aceste arii. O trăsătură

Info 7.1

Calificări pentru utilizarea testelor psihologice

Aşa cum am discutat în Capitolul 1, niciun set formal de atestate dobândite fie prin programe educaţionale, licenţe sau certificări, nu poate asigura competenţa utilizării unui anumit test într-o situaţie dată.
Calificările utilizatorilor testelor psihologice se bazează pe doi factori principali:

1. cunoştinţele şi abilităţile în
   - principii psihometrice şi statistică;
   - selectarea testelor prin prisma calităţilor lor tehnice şi interpretarea, raportarea şi protejarea rezultatelor lor;
   - toate aspectele relevante pentru contextul şi scopul utilizării testului.
2. măsura în care utilizatorii unui test au o experienţă adecvat supervizată în toate aspectele cunoştinţelor şi abilităţilor relevante pentru utilizarea specifică a unui anumit test.

Pentru mai multe informaţii, vezi Turner, S. M., DeMers, S. T., Fox, H. R. & Reed, G. M. (2001). APA's guidelines for test user qualifications: An executive summary. *American Psychologist*, 56, 1099–1113.

deosebit de utilă a documentaţiei realizate de U.S. Department of Education este că, pe lângă rezumarea principiilor de bază ale practicii de testare în domeniul evaluării educaţionale, aduce în discuţie, de asemenea, şi cerinţele legale care se aplică utilizării nediscriminatorii a testelor în scopuri deosebit de importante. Documentul SIOP, deşi nu intenţionează să interpreteze statute, reguli sau legi ce privesc deciziile de angajare, oferă principii pentru aplicarea şi utilizarea procedurilor de selecţie care pot informa şi ghida părţile responsabile de autorizarea şi implementarea acestor proceduri în evaluarea adecvării şi acceptabilităţii lor.

În ultimii ani, profesiile asociate cel mai strâns cu testarea psihologică — reprezentate în Joint Committee on Testing Practices (JCTP) — au ajuns să recunoască din ce în ce mai mult faptul că şi subiecţii testaţi trebuie să fie informaţi de drepturile şi responsabilităţile pe care le au în procesul testării. În acest scop, au publicat un document al cărui unic scop este de a oferi astfel de informaţii (JCTP, 1998). Info 7.2 enumeră unele dintre cele mai importante drepturi şi responsabilităţi care sunt subliniate şi descrise în acest document, disponibil în întregime în secţiunea Testing and Assessment a site-ului web APA (http://www.apa.org/science/ttrr.html).

Aplicarea testelor în scopuri inacceptabile sau în moduri neadecvate de către utilizatori cărora le lipsesc instruirea şi calificările necesare duce invariabil la utilizarea greşită a unui test.

Nu uitaţi!

- *Chiar şi instrumentele construite cu cea mai mare atenţie şi convingătoare din punct de vedere psihometric pot fi utilizate greşit.*
- Conform cu *Testing Standards* (AERA, APA, NCME, 1999, cap. 11), responsabilitatea pentru utilizarea adecvată a testelor şi interpretarea competentă a scorurilor la test îi revine în principal utilizatorului testului.
- Utilizarea greşită poate apărea în fiecare pas al procesului de testare, începând cu selectarea neadecvată a instrumentelor fie în raport cu scopul în care acestea sunt aplicate, fie în raport cu indivizii cărora le sunt administrate testele. Erorile din administrarea sau scorarea şi din interpretarea sau comunicarea rezultatelor unui test pot reprezenta o problemă de utilizare greşită a testelor.
- Ori de câte ori se intenţionează utilizarea testelor, cea mai bună cale de a preveni utilizarea lor greşită este de a ne asigura încă de la început că indivizii implicaţi în fiecare aspect al utilizării unui test au calificările şi competenţa necesare pentru a-şi îndeplini rolul în procesul testării.

Info 7.2

Drepturile și responsabilitățile subiecților testați

Documentul *Rights and Responsibilities of Test Takers: Guidelines and Expectations* (JCTP, 1998) poate fi reprodus și distribuit gratuit. Deși nu are forța unei legi — iar legile statale și federale se situează deasupra drepturilor și responsabilităților declarate aici —, profesioniștii implicați în procesul testării au responsabilitatea de a se asigura că subiecții testați sunt conștienți de informația conținută în document.

Printre *drepturile importante ale subiecților testați* se numără:

- Dreptul de a primi anterior testării o explicație despre (a) scopurile testării, (b) testele ce urmează a fi utilizate, (c) dacă rezultatele testelor le vor fi comunicate numai lor sau și altor persoane și (d) cum s-a planificat utilizarea rezultatelor. Dacă subiecții testați au o dizabilitate sau au dificultăți în înțelegerea limbajului testului, aceștia au dreptul să ceară sau să afle despre posibilitatea unor ajustări ale testării.
- Dreptul de a ști dacă testul este opțional și de a afla despre consecințele faptului de a se supune sau nu testării, de a completa testul sau de a anula scorurile.
- Dreptul de a primi o explicație a rezultatelor testului într-o perioadă de timp rezonabilă și într-un limbaj comun, ce poate fi înțeles.
- Dreptul la confidențialitatea rezultatelor testului în măsura permisă de lege.

Printre *responsabilitățile* importante ale persoanelor testate se numără:

- Responsabilitatea de a citi și/sau a asculta drepturile și responsabilitățile pe care le au.
- Responsabilitatea de a pune întrebări anterior testării cu privire la motivele pentru care se aplică testul, cum va fi aplicat, ce li se va cere să facă și ce se va întâmpla cu rezultatele.

Acest capitol se ocupă de câteva considerații esențiale de care trebuie să se țină cont în utilizarea testelor psihologice, inclusiv de aspectele legate de selectarea, administrarea și scorarea testelor, precum și de interpretarea și raportarea scorurilor unui test. Din moment ce mare parte din informația din capitolele anterioare este relevantă pentru aspectele fundamentale implicate în utilizarea responsabilă a testelor, cititorii pot să revină la părțile anterioare ale acestui volum atunci când în capitolul de față sunt menționate subiecte discutate anterior. În măsura în care utilizatorii de teste și subiecții testați sunt conștienți de ceea ce reprezintă o practică a testării competentă, posibilitatea de utilizare greșită a unui test scade și este mai probabil să se atingă beneficiile potențiale inerente în utilizarea testelor.

Aspecte esențiale ale selectării testelor

Așa cum am discutat în Capitolul 1, testele psihologice sunt utilizate în principal pentru a ajuta la luarea deciziilor cu privire la oameni în domeniile educațional, ocupațional, clinic, judiciar și în consilierea vocațională. În plus, testele sunt de asemenea utilizate frecvent în cercetarea psihologică și, recent, au început să fie aplicate în procesul de psihoterapie în scopul dezvoltării personale, al sporirii autocunoașterii sau în ambele (Finn & Tonsager, 1997). Consecințele utilizării greșite a testelor diferă larg în termenii impactului lor potențial asupra subiecților individuali testați. Bineînțeles, gradul de atenție acordată fiecărui pas al procesului testării — începând cu selectarea testului — trebuie să țină cont și să fie proporțional cu impactul pe care este probabil să îl aibă testarea. Deși cea mai mare parte a materialului care urmează se poate aplica și utilizării testelor în cercetare sau în

ATENȚIE!

Dintre toate ariile în care sunt utilizate testele psihologice, poate niciuna nu este mai controversată decât aplicarea lor în domeniul judiciar. Datorită naturii celor mai multe proceduri legale, profesioniștii care acceptă să apară ca experți în procese ce presupun dovezi obținute din testarea psihologică se pot aștepta ca bazele mărturiei lor să fie puse la îndoială și atacate cu orice ocazie. În astfel de situații, așa-numiții experți nepregătiți să-și apere mărturia bazată pe teste pot fi cu ușurință umiliți și înjosiți de avocații care sunt bine pregătiți pentru atac.

O cale instructivă pentru a vă familiariza cu potențialele capcane ale utilizării greșite a testelor în domeniul legal, precum și în alte contexte, este consultarea lucrării în mai multe volume *Coping With Psychiatric and Psychological Testimony*, scrisă de Jay Ziskin și colaboratorii săi și adresată în principal avocaților. A cincea și ultima ediție a acestei lucrări a fost publicată în 1995 de Law and Psychology Press din Los Angeles, CA. De atunci a fost îmbunătățită prin suplimentele publicate în 1997 și 2000. Dintre cele trei volume ale lucrării principale, volumul 2 — devotat în principal contestării mărturiilor derivate din testele psihologice de diferite tipuri — este cel mai relevant pentru subiectele luate în considerare în acest capitol.

scopuri terapeutice, punctul de plecare pentru prezentarea celor mai multe teme din acest capitol este presupunerea că testele sunt utilizate în principal pentru a lua decizii cu privire la oameni.

A folosi sau a nu folosi (teste), aceasta e întrebarea (prima)

Indiferent de scopurile urmărite de testarea psihologică, primul aspect care trebuie stabilit este dacă este nevoie de testare. Pentru a rezolva acest aspect, utilizatorii de teste ar trebui să se implice

într-o analiză costuri–beneficii — similară cu cea sugerată de Goldman (1971) — și să ia în considerare în mod explicit următoarele întrebări:

1. Ce tip de informație caut să obțin din testare?
2. Cum va fi utilizată acestă informație?
3. În ce măsură informația pe care o caut este deja disponibilă din alte surse?
4. Ce alte instrumente ar putea fi folosite pentru a obține informația pe care o caut?
5. Care sunt avantajele utilizării testelor în locul altor surse de informații sau alături de acestea?
6. Care sunt dezavantajele sau costurile în timp, efort și bani ale utilizării testelor în locul altor surse de informații sau alături de acestea?

Dacă argumentele pentru utilizarea testelor și aplicarea așteptată a rezultatelor unui test nu sunt explicite încă de la început, este probabil ca scorurile la test să nu fie de mare folos sau, mai rău, e probabil ca acestea să fie utilizate greșit. Dacă informația necesară este deja disponibilă din alte surse sau poate fi obținută prin alte mijloace, probabil că testarea va fi superfluă, cu excepția cazului în care scopul său este acela de a confirma ceva deja cunoscut sau de a aduce un sprijin suplimentar în acea direcție. În cele din urmă, dacă avantajele sau beneficiile care pot fi obținute prin utilizarea unui test nu depășesc în greutate costurile sau dezavantajele sale, inclusiv orice prejudiciu potențial care ar putea apărea din utilizarea unui test, testarea, în mod evident, nu e recomandabilă. Info 7.3 enumeră principalele motive și circumstanțe în care testarea poate fi nerecomandabilă.

Info 7.3.

Primele 10 motive pentru a *nu* utiliza teste

Din multe motive și în multe situații nu este recomandată utilizarea testelor psihologice; următoarea listă le prezintă pe cele mai importante. Cu câteva excepții, testele psihologice nu ar trebui utilizate în următoarele circumstanțe:

1. Scopul testării este necunoscut sau neclar pentru utilizatorul testului.
2. Utilizatorul testului nu este complet familiarizat cu toată documentația necesară și nu este instruit în procedurile legate de test.
3. Utilizatorul testului nu știe unde vor ajunge rezultatele testării, cum vor fi utilizate sau nu poate menține siguranța lor.
4. Informația urmărită prin testare este deja disponibilă sau poate fi obținută mai eficient din alte surse.
5. Subiectul testat nu dorește sau nu este capabil să coopereze în procesul testării.
6. Este probabil ca subiectul testat să fie vătămat datorită procesului de testare.
7. Cadrul și condițiile mediului de testare sunt neadecvate.
8. Formatul sau materialele testului sunt neadecvate în privința vârstei, sexului, bagajului cultural sau lingvistic, dizabilității sau oricărei alte condiții a subiectului testat care ar putea invalida datele testului.
9. Normele testului sunt învechite, inadecvate sau inaplicabile pentru subiectul testat.
10. Documentația despre fidelitatea sau validitatea scorurilor la test este neadecvată.

Două motive importante pentru utilizarea testelor

Deoarece testarea are loc în atât de multe cadre diferite și este utilizată în atât de multe scopuri diferite, este dificil să discutăm impactul ei la nivel abstract. Cu toate acestea, oricare ar fi contextul în care sunt utilizate testele psihologice — considerând că

au fost construite cu atenție și că este disponibilă documentația adecvată a valorii lor psihometrice pentru un scop dat — atunci când sunt utilizate în mod judicios, testele au avantaje distincte față de alte metode de colectare a informațiilor despre persoane. Cele mai semnificative avantaje pe care testele psihologice le oferă țin de eficiența și obiectivitatea lor caracteristică.

- *Eficiența.* Multor întrebări adresate prin testarea psihologică li se poate răspunde prin alte metode dacă indivizii care caută informațiile respective au timpul și resursele necesare pentru a le colecta. De exemplu, cei mai mulți oameni nu au nevoie de utilizarea testelor pentru a schița profilul psihologic al celor cu care au avut un contact personal extins și continuu. Prin observarea comportamentului indivizilor și interacțiunea cu aceștia în diferite situații pe o perioadă lungă de timp, pot fi acumulate date ample din care se pot trage concluzii cu privire la aptitudinile și atributele personale și chiar despre modul în care este probabil să se comporte în viitor. Testarea psihologică le permite celor care se află în poziția de a evalua și de a lua decizii cu privire la oameni — dar care nu au oportunitatea pentru observare și interacțiune de lungă durată cu aceștia — instrumentele pentru obținerea informațiilor de care au nevoie într-o manieră rapidă și eficientă.
- *Obiectivitate.* Chiar și atunci când există o oportunitate mare de a observa și a interacționa cu oamenii, datele pe care le putem acumula prin observații informale și interacțiuni pot fi puse la îndoială sau pot avea valoare limitată. Datele de observație obținute într-o manieră nesistematică pot conduce cu ușurință la judecăți imprecise din cauza subiectivității

observatorilor, incluzând variații în *ceea ce* observă și își amintesc, precum și în *cât de bine* observă, relatează și își evaluează observațiile. Nu e nevoie să spunem că această subiectivitate intră în joc în special atunci când observatorii nu sunt imparțiali, ca în cazul în care aceștia sunt prieteni, rude sau asociați ai persoanei observate. Chiar și atunci când observatorii sunt atenți și sunt capabili să se detașeze de subiectul observației lor, datele brute obținute din observații și interacțiuni naturale nu pot fi evaluate și interpretate în mod adecvat fără un anumit cadru de referință standardizat în raport cu care să fie comparate. Așa cum am discutat în capitolele anterioare, semnificația și valoarea ce poate fi obținută din eșantioanele de comportament obținute prin teste standardizate depind aproape în întregime de cadrele de referință normative sau bazate pe criteriu care sunt disponibile pentru comparații și de datele acumulate privind fidelitatea și validitatea scorurilor la test.

Utilitatea testelor

Ori de câte ori se are în vedere utilizarea testelor pentru a lua decizii cu privire la oameni, utilitatea potențială a folosirii lor trebuie să fie luată în considerare și analizată. Conceptul de *utilitate*, introdus succint în Capitolul 5 ca noțiune strâns legată de validitate, se referă la o evaluare a dezirabilității subiective a unui rezultat sau eveniment. Utilitatea este un aspect al temei mult mai largi a teoriei deciziei, o abordare larg utilizată de strângere și analizare a informațiilor, de obicei sub formă matematică, pentru a elabora strategii raționale de luare a deciziilor (vezi Bell, Raiffa, & Tversky, 1988). Estimarea valorii deciziilor corecte versus

costul deciziilor incorecte este, invariabil, o chestiune complicată, mai ales atunci când deciziile ce urmează să fie luate vor influența viețile oamenilor în moduri semnificative. Teoria deciziilor se ocupă de dezvoltarea și analiza strategiilor posibile de luare a deciziilor, utilizând informația disponibilă pentru a estima utilitatea lor prin calcularea costurilor și beneficiilor diferitelor rezultate în termeni cantitativi — de obicei monetari. Bineînțeles, pentru a estima utilitatea unui rezultat, trebuie adoptat un punct de vedere subiectiv care derivă, de obicei, din valorile și prioritățile celor care iau decizia.

De exemplu, în domeniul selecției de personal există două decizii posibile (angajare vs. neangajare). Presupunând că se poate dihotomiza criteriul performanței în muncă în categoriile succes și eșec, aceste decizii au patru rezultate posibile: (a) acceptare validă, (b) respingere validă, (c) acceptare falsă, (d) respingere falsă. Acceptările valide sunt cele mai dezirabile dintre toate rezultatele; nu numai că nu presupun niciun hazard pentru angajator sau candidatul angajat, dar, de fapt, oferă beneficii amândurora. Acceptările false, pe de altă parte, presupun anumite riscuri pentru angajator (ex.: pierderea unor profituri sau daune apărute din cauza incompetenței angajaților, timp și efort pierdute în activitățile de angajare și de instruire, o rată sporită a schimbării colectivului de muncă), dar, de obicei, nu și pentru angajat. Respingerile valide sunt în cea mai mare parte avantajoase pentru angajator, dar, cel mai probabil nu este la fel și pentru candidații respinși. Respingerile false, în mod tipic, nu prezintă riscuri semnificative pentru angajat, în afară de cazul în care deciziile sunt contestate, dar pot dăuna semnificativ candidaților respinși în mod fals.

Este clar că utilitatea rezultatelor diferă pentru diferitele părți afectate de decizie. Mai mult, consecințele potențiale ale deciziei

pentru toate părțile implicate nu pot fi niciodată anticipate în totalitate, cu atât mai mult cuantificate. Însă ideea de a încerca în mod explicit anticiparea și evaluarea beneficiilor și riscurilor probabile inerente în utilizarea testelor — în opoziție cu alte instrumente — se află în centrul utilizării responsabile a testelor. În acest scop, teoria deciziei oferă câteva concepte fundamentale care pot fi utilizate atunci când se ia în considerare dacă și în ce măsură testele pot contribui la rezultate îmbunătățite prin creșterea numărului de decizii corecte și minimizarea numărului de decizii incorecte. Aceste concepte sunt potrivite mai ales pentru organizații care caută să determine posibilele câștiguri în corectitudinea deciziilor de selecție ce pot fi obținute prin implementarea unei strategii de luare a deciziilor ce include utilizarea scorurilor obținute la un test (vezi Kobrin, Camara, & Milewski, 2002). Următoarele elemente informaționale sunt esențiale pentru estimarea acestor câștiguri:

- *Datele validității.* Scorurile la test pot îmbunătăți luarea deciziilor doar în măsura în care prezintă validitate demonstrabilă a evaluării poziției subiecților testați sau a predicției performanțelor acestora la un criteriu; cu cât este mai mare coeficientul de validitate al scorurilor la test, cu atât mai mare va fi precizia estimărilor criteriului și a predicțiilor (vezi Capitolul 5).
- *Datele ratei de bază:* conceptul de *rată de bază* se referă la punctul de pornire în raport cu care pot fi calculate câștigurile potențiale în precizie, adică probabilitatea apariției anumitor evenimente într-o populație anterior introducerii oricărei proceduri noi, cum ar fi testarea. De exemplu, dacă proporția de studenți care încheie un program de studii este

0,60 sau 60% din cei admiși în program, rata de bază este 0,60. În deciziile de selecție de personal, rata de bază este proporția de candidați angajați pentru o anumită poziție care prezintă succes la locul de muncă. În scopul calculării *validității incrementale* sau a îmbunătățirilor aduse de un test în deciziile de selecție, rata de bază se referă la proporția de decizii corecte luate fără utilizarea scorurilor obținute la test. Cele mai mari câștiguri potențiale în precizia deciziilor de selecție apar atunci când rata de bază este apropiată de 0,50. Rate de bază apropiate de extreme (e.g. 0,10 sau 0,90) indică faptul că o selecție precisă este fie foarte dificilă, fie foarte ușoară în circumstanțele date. În astfel de cazuri, utilizarea scorurilor la test ca bază pentru selecție — chiar dacă au un grad relativ înalt de validitate — poate să nu sporească precizia deciziilor, și chiar o poate scădea.

- *Rate de selecție:* o altă restricție a potențialelor contribuții pe care scorurile obținute la un test le pot avea la îmbunătățirea deciziilor de selecție apare din *rata de selecție*, care este rata ce rezultă atunci când numărul de posturi libere este împărțit la numărul de candidați pentru acele poziții. Dacă sunt libere 5 posturi pentru care candidează 500 de oameni, rata de selecție este 5 ÷ 500 sau 1%, în timp ce, dacă există doar 25 de candidați, rata de selecție este 5 ÷ 25 sau 20%. Bineînțeles, rate de selecție mai mici oferă celor care iau deciziile oportunitatea de a fi mai selectivi decât în cazul ratelor mari. Ca un exemplu extrem, dacă 3 poziții sunt libere și candidează 3 persoane, rata de selecție este 100% și angajatorul nu dispune de nicio alegere dacă organizația are nevoie de acești lucrători. Într-o astfel de situație, validitatea incrementală pe care scorurile la test o pot avea este

nulă — indiferent de cât de valide pot fi din punct de vedere predictiv. Pe de altă parte, atunci când rata de selecție este foarte redusă, chiar și un test ale cărui scoruri au doar un grad moderat de validitate predictivă poate ajuta la creșterea ratei deciziilor precise.

Așa cum a fost clar în urma discuțiilor anterioare, gradul de îmbunătățire a preciziei deciziilor de selecție care poate fi obținut prin utilizarea scorurilor la test în cadrul ocupațional și în cel educațional depinde de o combinație a ratelor de bază și a ratelor de selecție într-o situație dată, precum și de validitatea scorurilor obținute la test. Efectele interactive ale acestor trei variabile asupra preciziei selecției au fost luate în considerare de mult timp; de fapt, încă din 1939, Taylor și Russell au publicat un set de tabele care prezintă proporția așteptată a succeselor (acceptări valide) care sunt de așteptat în cazul diferitelor combinații de rate de bază, rate de selecție și coeficienți de validitate. Tabelele Taylor-Russell, care au precedat apariția teoriei deciziei, oferă o cantitate fundamentală de informații care pot fi folosite pentru a evalua posibilitatea de a utiliza un test dat într-un anumit context, dar ele nu evaluează toate rezultatele posibile ale unei decizii și nici nu aduc în discuție toți factorii relevanți pentru deciziile de selecție.

De la publicarea tabelelor Taylor-Russell, psihologii din cadrul industrial-organizațional au făcut progrese și au dezvoltat alte moduri de a evalua efectele utilizării testelor, încorporând conceptele și modelele teoriei deciziilor. Printre modificările apărute se numără estimarea efectelor utilizării testelor asupra altor rezultate decât acceptările valide, cum ar fi respingerile false, și asupra rezultatelor determinate prin criterii continue sau gradate care pot fi mai realiste pentru aplicațiile din luarea deciziilor decât

simpla dihotomie succes-eșec prezentă în tabelele Taylor-Russell. Aspecte suplimentare ale luării deciziilor, cum ar fi utilizarea predictorilor multipli în diferite combinații și evaluarea strategiilor de selecție, clasificare și repartizare a personalului în termenii utilității lor din diferite puncte de vedere (ex.: creșterea productivității, descreșterea schimbărilor colectivului de muncă, minimizarea costurilor de instruire etc.) au fost investigate în raport cu utilizarea testelor în managementul resurselor umane (vezi Boudreau, 1991; Schmidt & Hunter, 1998). Cititorilor interesați de o descriere detaliată a uneia dintre cele mai complete aplicații a numeroaselor concepte din teoria deciziilor la construirea și evaluarea unei baterii de teste și alte instrumente pentru selecția și clasificarea personalului li se recomandă prezentarea Proiectului A al U.S. Army făcută de J. P. Campbell și Knapp (2001) (vezi Capitolul 5 pentru mai multe informații despre acest subiect).

Utilitatea testelor în luarea deciziilor clinice

Unele dintre conceptele teoriei deciziilor pot fi aplicate pentru a maximiza eficiența testării și evaluării în alte arii decât selecția de personal. În psihologia clinică, de exemplu, *rata de bază* se referă la frecvențele cu care condițiile patologice, cum ar fi depresia, apar într-o populație dată. Dat fiind că informația necesară este disponibilă, cunoașterea ratei de bază poate ajuta la evaluarea posibilei utilități a folosirii testelor în scop diagnostic sau predictiv în cadrul unei populații date. Așa cum de întâmplă și în cazul deciziilor de selecție, cea mai mare contribuție pe care testele o pot avea la îmbunătățirea preciziei predictive sau diagnostice apare atunci când ratele de bază sunt mai apropiate de 0,50. Atunci când o condiție ce urmează a fi diagnosticată este fie pervazivă, fie foarte rară în cadrul unei populații, adică ratele de bază sunt fie foarte

ridicate, fie foarte scăzute, îmbunătățirile preciziei diagnostice obținute prin utilizarea testelor diagnostice sunt mai scăzute decât atunci când tulburarea respectivă apare cu o frecvență moderată în populație. Ca și în cazul deciziilor de selecție, motivul este dat de faptul că ratele de bază au un efect restrictiv asupra preciziei deciziilor. Cu privire la diagnostic, dacă rata de bază a unei tulburări din cadrul unei populații este foarte ridicată (e.g. 0,80), probabilitatea unui *rezultat pozitiv fals* — adică diagnosticarea unei tulburări atunci când nu este prezentă — este scăzută (0,20), chiar dacă deciziile de diagnostic sunt făcute pe o bază aleatorie. În mod similar, dacă rata de bază este extrem de scăzută, probabilitatea unui *rezultat negativ fals* — adică nedetectarea condiției atunci când ea este prezentă — este de asemenea redusă. În astfel de situații, costul utilizării scorurilor la test pentru a identifica prezența unei condiții în cadrul unui întreg grup sau al unei populații poate să nu fie justificată de câștigurile de precizie relativ modeste pe care le pot aduce scorurile la test. Pentru o prezentare excelentă a implicațiilor ratelor de bază în clinică, i se recomandă cititorului capitolul scris de Finn și Kamphuis (1995) pe această temă.

Trebuie observat că în cadrul clinic, ca și în alte domenii ale practicii evaluării psihologice, deciziile sunt luate de obicei pe bază individuală, utilizând datele obținute prin intermediul unor instrumente diferite, inclusiv cunoștințele de specialitate și judecata evaluatorului. Trecerea în revistă a tuturor înregistrărilor medicale, inclusiv a rezultatelor testelor de laborator și a rezultatelor obținute prin diferite tehnici imagistice disponibile în mod curent pentru a explora sau a exclude factori fiziologici (ex.: consumul sau consumul excesiv de droguri, tulburări neurologice sau endocrine etc.) ce pot cauza sau pot contribui la simptome psihice este o parte indispensabilă a evaluărilor psihiatrice și

neuropsihologice în cadre clinice. Prin urmare, contribuția adusă de rezultatele testelor psihologice în astfel de cadre nu poate fi evaluată izolat. În plus, utilitatea potențială a deciziilor de diagnostic individual bazate doar pe rezultatele la test este redusă de ratele de bază scăzute ale celor mai multe tulburări mentale în populația generală (O'Leary & Norcross, 1998), de coeficienții de validitate relativ mici ai multor teste utilizate în astfel de scopuri și de natura evolutivă a nosologiei psihiatrice și a criteriilor de diagnostic, mai ales cu privire la tulburările de personalitate (vezi Livesley, 2001). Mai mult, aspectele contextuale legate de motivele pentru care este realizată evaluarea diagnostică pot influența semnificativ relativa ușurință sau dificultate a obținerii unor rezultate precise prin intermediul testării psihologice sau al altor instrumente de evaluare. Din perspectiva tuturor acestor factori, nu este surprinzător că utilitatea testelor în luarea deciziilor clinice este subiectul unei dezbateri frecvente în domeniul psihologiei. Info 7.4 enumeră câteva dintre cele mai notabile contribuții ale diferitelor părți implicate în această dezbatere.

Alte instrumente de evaluare

Există, bineînțeles, multe modalități ce pot fi utilizate în locul testelor sau alături de teste pentru a obține informații necesare pentru evaluarea psihologică și luarea deciziilor cu privire la oameni. Printre instrumentele utilizate cel mai frecvent în evaluarea indivizilor se numără istoriile de caz sau datele biografice, interviurile, observația sistematică sau cea desfășurată în cadru natural, înregistrările academice sau de angajare precum și referințele din partea profesorilor și supervizorilor. Fiecare dintre aceste surse de informație, singură sau în combinație cu altele,

Info 7.4

Sunt utile testele psihologice în luarea deciziilor clinice?

Această întrebare este subiectul unei dezbateri aproape la fel de ample ca și întrebarea dacă psihoterapia este eficientă în tratarea tulburărilor mentale. Puțini ar răspunde la aceste întrebări cu un da sau un nu definitiv. Cei mai mulți practicieni care utilizează teste în cadrul clinic, în consiliere sau în cadrul judiciar cred cu tărie în valoarea lor, în timp ce deseori opusul este adevărat pentru nepracticieni. Pentru diferite opinii despre acest subiect, cititorii pot consulta următoarele surse:

- Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T.W., Meyer, G. J. & Moreland, K. (2000). Problems and limitations in the use of psychological assessment in the contemporary health care delivery system. *Professional Psychology: Research and Practice, 31,* 131–140.
- Hummel, T. J. (1999).The usefulness of tests in clinical decisions. În J.W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 59–112). Boston: Allyn & Bacon.
- Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., & Eisman, E. J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice, 31,* 119–130.
- Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165. (Vezi secțiunea Comentarii a numărului din februarie 2002 al revistei *American Psychologist* pentru răspunsuri la articolul scris de Meyer et al.).

poate furniza date valoroase și poate contribui la reușita în luarea deciziilor. De fapt, câteva dintre aceste instrumente de evaluare (interviul, evaluările observatorilor) pot fi standardizate și evaluate în termenii fidelității și validității lor și din perspectiva

utilității lor în luarea deciziilor. Atunci când se fac afirmații cu privire la validitatea unor astfel de instrumente pentru un scop specific, utilizarea lor intră în aria indicațiilor din *Testing Standards* (AERA, APA, NCME, 1999, pp. 3–4). Dacă datele pe care aceste metode le oferă sunt fidele și valide în mod demonstrabil, ele pot fi suficiente ca bază pentru luarea deciziilor, fie singure, fie în combinație cu altele.

Date biografice

Informațiile despre istoria de viață, cunoscute și ca *date biografice*, pot fi obținute prin numeroase metode, inclusiv prin interviuri, chestionare și examinarea înregistrărilor existente ale comportamentului trecut, cum ar fi foi matricole, rapoarte ale poliției și așa mai departe. Așa cum am afirmat mai devreme, examinarea înregistrărilor din dosarele medicale este un aspect indispensabil al oricărei evaluări clinice a simptomelor care pot să provină din tulburări neurologice și metabolice, ingerare de medicamente sau alte condiții fizice sau să fie influențate de acestea. Mai mult, în aproape orice tip de evaluare, datele bine documentate despre comportamentul trecut și despre realizările unei persoane se numără printre cele mai valide și mai fidele surse de informații evaluative deoarece sunt dovezi factuale pentru ceea ce un individ este capabil să facă sau să realizeze. De exemplu, cu privire la întrebarea dificilă a predicției comportamentului violent, abordarea actuarială a evaluării riscului oferă predicțiilor un punct de ancoră fundamental. Această abordare — care poate fi suplimentată de o evaluare a statutului curent, precum și de informații clinice și din testare — se bazează pe examinarea sistematică a datelor istoriei personale care au demonstrat relații empirice cu periculozitatea, cum ar fi violența anterioară, consumul excesiv

de substanțe, stabilitatea locului de muncă, tulburările mentale și neadaptarea timpurie (Borum, 1996; Monahan & Steadman, 2001). Un exemplu similar al valorii datelor istorice poate fi extras din domeniul ocupațional, unde, în general, se consideră că o înregistrare demonstrabilă a succeselor trecute într-o ocupație este unul dintre cei mai buni determinanți ai succeselor viitoare. Astfel, dacă informația despre o performanță trecută în muncă a unui potențial angajat este curentă, fidelă și înalt favorabilă, cei mai mulți angajatori nu vor ezita să angajeze acel individ pe această bază. O abordare mai structurată și formală a utilizării informației despre istoria de viață a unei persoane este utilizarea datelor obținute prin intermediul inventarelor biografice standardizate, care pot fi validate ca predictori ai performanțelor viitoare în diferite contexte. Pentru o descriere completă a utilizării informațiilor biografice în luarea deciziilor de selecție, cititorii interesați ar putea consulta *Biodata Handbook*, editată de Stokes, Mumford și Owens (1994).

Datele de interviu

Interviul poate oferi informații bogate în aproape orice context de evaluare, permite ocazia de a observa comportamentul intervievatului și de a colecta date relevante despre istoria de viață și informații cu privire la atitudinile, opiniile și valorile individului. Interviul față în față este un instrument foarte flexibil și, atunci când este condus adecvat, se poate dovedi de importanță critică în luarea deciziilor cu privire la oameni. În multe cadre clinice, judiciare și ocupaționale, intervievarea indivizilor evaluați și a celor care pot oferi date colaterale este considerată un aspect esențial al procesului de evaluare (vezi Hersen & Van Hasselt, 1998). Cu toate acestea, fidelitatea și validitatea datelor

interviului sunt puternic dependente de abilitățile intervievatorului și de obiectivitatea sa în colectarea, înregistrarea și interpretarea informațiilor. Pentru a combate punctele slabe potențiale inerente în datele de interviu, practicile curente din cele mai multe domenii care utilizează tehnici de interviu subliniază fie instruirea intensivă a intervievatorilor, fie utilizarea interviurilor structurate — care sunt instrumente cu adevărat standardizate, asemănătoare cu testele în multe moduri — sau ambele posibilități.

Observația

O altă sursă de date, frecvent întâlnită, constă în evaluări, liste și sisteme de cotare bazate pe diferite tipuri de observații directe ale comportamentului sau pe contacte anterioare relevante cu persoana care urmează să fie evaluată. Ca și în cazul intervievării, fidelitatea și validitatea evaluărilor și a altor date derivate din observație pot varia mult în funcție de evaluator și de sistemul utilizat pentru evaluarea sau înregistrarea observațiilor. Instruirea oferă evaluatorilor un standard uniform cu ajutorul căruia să evalueze performanța și poate îmbunătăți calitatea datelor de observație. Utilizarea scalelor de evaluare standardizate pentru a colecta datele de la persoane care dețin informații, cum sunt părinții sau profesorii, este o procedură obișnuită în evaluarea copiilor și adolescenților (vezi Kamphaus & Frick, 2002, cap. 7 și 8).

Surse suplimentare de informație

Multe dintre sursele alternative de informații utilizate în evaluarea indivizilor nu pot fi evaluate în mod adecvat cu privire la fidelitate și validitate. De exemplu, valoarea scrisorilor de

recomandare și a referințelor similare depinde de măsura în care persoanele care le oferă sunt dispuse și capabile să dezvăluie informații relevante într-o manieră corectă și completă. Cei care primesc referințele deseori nu cunosc factorii care ar putea să influențeze adevărul acestora. În mod similar, în ultimele decenii, numărul mare de calificative pentru toate nivelurile de școlarizare a erodat semnificația multor credite academice până în punctul în care nota medie și chiar anumite calificări și diplome nu mai pot fi considerate ca indicând exact anumite competențe, așa cum ar părea la prima vedere.

Căutarea testelor și evaluarea lor

Atunci când se urmărește utilizarea testelor psihologice într-un anumit scop, utilizatorii de teste se confruntă cu două sarcini majore și anume (a) descoperirea instrumentelor disponibile pentru scopurile pe care le au în vedere și (b) evaluarea lor din acea perspectivă și din punctul de vedere al persoanelor care urmează să fie testate. Cu privire la prima dintre aceste sarcini, li se recomandă cititorilor secțiunea despre Surse de informații despre teste din Capitolul 1 al acestei cărți și, în special, materialul din *FAQ/ Finding Information About Psychological Tests* (APA, 2003), creat și menținut de colectivul Departamentului de testare și evaluare psihologică din APA și disponibil pe site-ul web APA. Așa cum am menționat în Capitolul 1, această resursă este un punct de plecare excelent pentru persoana care caută informații despre testele psihologice publicate sau nepublicate. Documentul FAQ enumeră și rezumă conținuturile tuturor lucrărilor de referință despre testele de ambele tipuri și oferă informații despre modul în care pot fi localizate testele și editurile care publică teste, despre

modul în care pot fi cumpărate testele, despre softurile specifice, serviciile de scorare și informații suplimentare despre utilizarea adecvată a testelor.

Evaluarea instrumentelor identificate ca fiind posibil utile pentru un anumit scop constituie o problemă mai complexă. Cu siguranță, recenziile și lucrările despre testele în chestiune oferă în mod tipic informații de bază despre tipurile de scoruri și acoperirea oferită de un test, scopul și populația căreia îi este destinat, modul în care a fost construit, procedurile de standardizare, tipurile de itemi pe care îi utilizează, sursele posibile de aspecte preferențiale și caracteristicile practice ale conținutului

Nu uitați!

Articolul lui Lawrence Rudner (1994) despre „Întrebări pe care trebuie să le punem când evaluăm testele" diponibil pe internet la http://edres.org/pare, oferă un bun punct de plecare pentru evaluarea testelor. Cele mai multe dintre capitolele anterioare din această carte conțin un material relevant pentru evaluarea testelor. O scurtă trecere în revistă a anumitor aspecte cheie care trebuie luate în considerare în procesul de selectare a testelor poate fi găsită în:

- Info 3.2: Informații necesare pentru a evalua aplicabilitatea unui eșantion normativ
- Info 4.7: Considerații legate de fidelitate în selectarea testelor
- Info 5.8: Relația dintre întrebări, decizii și predicții care necesită validare în raport cu criteriul
- Info 5.12: Strategii de validare în raport cu interpretarea scorurilor la test
- Tabelul 5.1: Aspecte ale validității de construct și a surselor de dovezi legate de aceasta
- Info 6.6: De ce eșantioanele de comportament obținute prin itemii unui test sunt atât de complexe?

și formatului testului care influențează ușurința cu care poate fi administrat și scorat. Aceste informații îl pot ajuta pe utilizatorul testului să decidă dacă instrumentul, în forma actuală, pare a fi adecvat scopului și persoanelor cărora le este destinat sau dacă trebuie colectate norme locale sau date suplimentare despre fidelitate și validitate.

Cu toate acestea, chiar și atunci când un test este considerat ca fiind adecvat din punct de vedere psihometric, utilizatorul testului se confruntă cu multe alte aspecte practice care țin de contextul și situația specifică a persoanei testate, de nivelul de calificare și experiență pe care-l necesită utilizarea unui test, constrângerile de timp și cele financiare și așa mai departe. Info 7.5 enumeră câteva dintre variabilele care trebuie luate în considerare cu privire la indivizii sau grupurile specifice cărora le va fi administrat testul.

Dacă investiția de timp, efort și resurse financiare necesară pentru a implementa utilizarea unui test într-un context dat pare să fie justificată de informația pe care scorurile testului o pot furniza, următorul pas pentru viitorii utilizatori de teste este de a obține testul și de a îl încerca pe ei înșiși sau pe unul sau mai mulți indivizi dornici să se supună unei administrări practice a testului. În acest scop, multe edituri pun la dispoziție *seturi de specimene* ale instrumentelor care conțin manualul testului și exemple de materiale necesare pentru administrare și scorare. Studierea atentă a acestor materiale, în special a manualului testului, este considerată de *Testing Standards* drept o condiție preliminară a utilizării reușite a unui test, pentru că oferă documentația necesară utilizatorilor testului pentru a evalua măsura în care rezultatele testului vor servi scopurilor în care intenționează să utilizeze testul. Mai mult, documentația testului oferită

Info 7.5

Câteva variabile legate de persoanele testate care trebuie luate în considerare în selectarea testelor

- *Variabile legate de mediul de testare*: alegerea unui mediu specific de prezentare a itemilor testului, cum ar fi creion-hârtie versus administrare pe computer sau prezentare orală versus vizuală va aduce examinaților anumite avantaje sau dezavantaje, în funcție de familiaritatea lor cu mediul, de acuitatea senzorială vizuală sau auditivă, de abilitățile lor motorii și așa mai departe. Chiar și atunci când un test include anumiți itemi de exersare pentru a familiariza subiecții testați cu mediul ales, este probabil ca diferitele lor grade de experiență anterioară cu acel mediu să continue să influențeze performanța atinsă la test.
- *Variabile legate de formatul testului*: Indiferent de conținutul lor, itemii cu răspuns la alegere tind să necesite mai multe abilități receptive decât itemii cu răspuns construit, care presupun utilizarea abilităților expresive. Din moment ce subiecții cărora li se administrează testul diferă în privința acestor abilități, alegerea formatului itemului va afecta și performanța lor. Variabile suplimentare, cum ar fi utilizarea limitelor de timp sau administrare individuală versus administrare în grup, vor afecta, de asemenea, în mod diferențial subiecții testați pe baza bagajului lor cultural și experiențial.
- *Variabile legate de limbajul itemilor testului:* Ori de câte ori limbajul este o parte a unui test, dar nu esența acestuia, abilitățile lingvistice receptive și expresive ale subiecților testați pot influența scorurile în mod neadecvat. Astfel, în funcție de cerințele unui anumit test, este necesar ca examinatorii să ia în considerare și să stabilească dacă subiecții testați au un nivel al vocabularului suficient și dețin abilitățile de lectură și de scriere pentru a îndeplini sarcinile care li se cer. În acest scop, de exemplu, manualele testelor trebuie să includă informații cu privire la nivelul de lectură necesar pentru a înțelege itemii testului și deseori se întâmplă astfel.

> **Nu uitați!**
>
> O resursă unică pentru exersarea citirii și interpretării materialului din manualul testului este furnizată în cartea lui Ann Corwin Silverlake (1999) *Comprehending Test Manuals: A Guide and Workbook*, publicată în Los Angeles de Pyrczak.

de editurile care publică teste trebuie să specifice calificările necesare pentru administrarea și interpretarea precisă a scorurilor obținute la test și, ori de câte ori este posibil, trebuie să avertizeze cititorii cu privire la posibila utilizare greșită a unui test (AERA, APA, NCME, 1999, cap. 5). Astfel de documentații pot cuprinde manuale tehnice, indicații pentru utilizator și alte materiale similar destinate să suplimenteze manualul testului.

Aspecte esențiale ale administrării și scorării unui test

Spre deosebire de alte aspecte ale utilizării testelor luate în considerare în capitolul de față, esențialul administrării unui test poate fi rezumat cu ușurință în două cuvinte: pregătire adecvată. Administrarea adecvată a testelor psihologice necesită o pregătire atentă a mediului de testare, a subiectului testat și a persoanei care administrează testul.

Pregătirea mediului de testare

Cele mai importante principii care trebuie urmate în pregătirea mediului în care va avea loc testarea sunt anticiparea și înlăturarea oricărei surse potențiale de distragere a atenției. Camerele în care are loc testarea trebuie să fie luminate și ventilate adecvat, trebuie să ofere scaune și spații adecvate pentru cei testați și nu

trebuie să existe zgomote sau alți stimuli (ex.: mâncare sau băutură) care pot să influențeze capacitatea subiecților testați de a se ocupa de sarcinile lor. Pentru a preveni întreruperile posibile, de obicei se așază un semn — pe care multe edituri îl oferă gratuit — pe ușa sălii de examinare pentru a avertiza trecătorii că în acel loc se desfășoară o testare.

După asigurarea unei săli de testare potrivite, pregătirea adecvată a mediului de testare presupune urmarea instrucțiunilor de administrare din manualul testului, care urmăresc repetarea cât mai bună a condițiilor în care testul a fost standardizat. În cazul testării de grup, aceste instrucțiuni pot include oferirea materialelor necesare pentru a susține testul (creioane cu gumă) și asigurarea faptului că subiecții cărora li se aplică testul sunt așezați astfel încât nu pot discuta între ei sau nu pot privi în foaia de răspuns a celorlalți, sunt angajați supraveghetori și așa mai departe. Pentru testele individuale, sunt necesare aranjamente speciale de testare, astfel încât examinatorul să poată prezenta materialele testului în direcția adecvată, să înregistreze răspunsurile fără să atragă atenția și să observe în totalitate comportamentul subiectului testat. Ca regulă generală, nu trebuie permisă prezența oricărei alte persoane în afară de examinator și subiectul testat în camera în care are loc administrarea individuală a testului (vezi National Academy of Neuropsychology, 1999). Prezența unei a treia părți aduce posibilitatea de distragere a atenției și chiar de influențare a procesului de testare, introduce un element neconcordant cu administrarea standardizată a testului și un risc suplimentar și inutil pentru siguranța testului. Pot exista circumstanțe speciale care cer observarea administrării testului de către alte persoane — de exemplu, studenți aflați în instruirea formală a administrării testului. În mod ideal, astfel de observații

trebuie făcute dintr-o cameră alăturată cu sala de testare, printr-o oglindă unidirecțională. Pentru persoanele testate care ar putea avea dificultăți în comunicarea cu un examinator datorită vârstei mici, a nestăpânirii limbii în care se desfășoară testarea sau datorită unei dizabilități, poate fi necesară prezența unui părinte sau a unui interpret în camera de testare. În astfel de situații, ca și în orice alt caz în care sunt făcute aranjamente speciale care pot influența interpretarea scorurilor, acestea trebuie să fie incluse în raportul rezultatelor la test.

Pregătirea persoanei testate

Există două aspecte distincte ale testării legate de pregătirea persoanelor testate. Primul se află în general în raza de influență a examinatorului și privește stabilirea raportului cu persoana testată și oferirea instrucțiunilor adecvate înainte de administrarea testului. Al doilea aspect, care nu se află sub controlul examinatorului, ține de toate experiențele de viață anterioare pe care subiectul testat le-a putut avea și care i-ar putea influența performanța la un anumit test.

Stabilirea raportului

În contextul testării, termenul *raport* se referă la relația armonioasă care ar trebui să existe între persoana testată și examinator. Pentru a maximiza fidelitatea și validitatea rezultatelor la test, trebuie stabilită o atmosferă prietenoasă încă de la începutul ședinței de testare și, ideal, raportul ar trebui să se încadreze între bun și excelent. Bineînțeles, absența acestui raport poate fi atribuită fiecărei părți implicate în situația de testare și poate apărea din lipsa de experiență sau incompetența examinatorului,

o dispoziție nefavorabilă din partea persoanei testate sau ambele. Nu este nevoie să spunem, în măsura în care lipsește raportul bun, că performanța la test este probabil să fie influențată negativ, chiar până în punctul în care scorurile obținute la test sunt invalidate. Pentru a construi un raport bun, examinatorul ar trebui să încerce să implice în procesul de testare interesul și cooperarea persoanelor testate, astfel încât acestea să poată reacționa la sarcinile testului într-o manieră adecvată, depunând cele mai mari eforturi la testele de aptitudine și răspunzând cât mai deschis și onest posibil la instrumente construite pentru evaluarea personalității.

Procesul de construire a raportului este mai extins în cazul testării individuale decât în testarea de grup, deoarece testarea individuală permite examinatorilor să observe comportamentul persoanei testate îndeaproape și în mod continuu și să-și extindă eforturile de a menține raportul pe toată durata procesului de testare. Cu toate acestea, chiar și în cazul testării de grup, examinatorii trebuie să încerce să explice scopul testării, procedurile de testare și așa mai departe (vezi Info 7.2) într-o manieră prietenoasă, respectând restricțiile indicațiilor pentru administrarea testului oferite în manualul testului, care trebuie să fie urmate pentru a menține condiții uniforme.

Pregătirea testării din punctul de vedere al persoanei testate

Ar lipsi ceva dacă nu am spune că scopul în care are loc testarea și aspectele contextuale ale situației și circumstanțelor în care are loc testarea au o influență semnificativă asupra atitudinii și motivației subiecților testați. Dintre numeroși factori emoționali și cognitivi care au o influență asupra predispozițiilor examinaților față de testare și a măsurii în care aceștia sunt pregătiți pentru o

experiență de testare, anxietatea de testare și gradul de sofisticare a subiectului sunt probabil elementele cel mai frecvent discutate și investigate în literatura de specialitate. În continuare va fi prezentată o scurtă trecere în revistă a lor.

Anxietatea de testare. Perspectiva de a fi evaluat tinde să trezească un anumit grad de neliniște pentru toți subiecții testați, mai ales atunci când scorurile obținute la test urmează să fie utilizate drept bază pentru luarea unor decizii cu consecințe importante pentru ei. În unele cazuri, această neliniște este disipată cu ușurință și indivizii îi pot face față — și chiar poate îmbunătăți nivelele de performanță prin creșterea excitației fiziologice. În altele, anxietatea de testare devine un factor care scade nivelul de performanță al indivizilor sau chiar îi aduce în stare de incapacitate, ceea ce poate avea un efect dăunător asupra performanței la test sau îi poate împiedica să participe la testare. Mai mult, motivele pentru care persoanele testate trăiesc anxietate înainte și în timpul procesului de testare pot, de asemenea, varia mult. Unele dintre aceste motive pot fi legate de tipul de test pe care urmează să-l susțină (teste de rapiditate, teste de matematică etc.), altele pot fi legate de subiecții testați (așteptări negative bazate pe experiența anterioară), în timp ce altele pot depinde de contextul în care are loc testarea (e.g. testarea pentru angajare) sau de o combinație de variabile. Deși examinatorii ar trebui să fie alerți și să încerce să reducă nivelele de anxietate ale subiecților testați ca parte a procesului de construire a relației, există și măsuri pe care subiecții testați le pot lua în același scop. Info 7.6 enumeră resursele care pot fi utile pentru subiecții testați și pentru cei interesați în abordările actuale ale evaluării și tratamentului anxietății de testare.

Sofisticarea. Strict vorbind, variabila cunoscută drept *gradul de sofisticare a subiectului* (numită și abilitatea de a fi testat sau înțelepciunea în raport cu testul) se referă la măsura în care subiecții testați au experiență sau exercițiu de testare. Ca regulă generală, pentru cele mai multe tipuri de teste de aptitudine, o experiență

Info 7.6

Surse de informații despre anxietatea de testare

Subiecții testați care vor ajutor pentru a gestiona anxietatea de testare vor găsi multe materiale în librării și pe internet. Printre acestea se numără:

- *Taking the anxiety out of taking tests: A step-by-step guide,* de S. Johnson. New York: Barnes & Noble Books, 2000.
- *No more test anxiety: Effective steps for taking tests and achieving better grades* de E. Newman (însoțită de CD audio). Los Angeles: Learning Skills Publications, 1996.
- The Test Anxiety Scale (Saranson, 1980) oferă o cale rapidă de a determina măsura în care o persoană poate avea tendința de a trăi anxietate de testare și este disponibilă gratuit la Learning Skills Publications (http://www.learningskills.com/test.html) și pe alte câteva site-uri de pe internet.
- Multe site-uri Web sponsorizate de centrele universitare de consiliere sunt accesibile căutând după criteriul „test anxiety" pe internet; aceste site-uri oferă informații utile despre obiceiurile de învățare și alte informații despre gestionarea anxietății de testare.

Celor care doresc să afle mai mult despre teoriile și cercetările actuale despre anxietatea de testare, le sunt recomandate următoarele lucrări:

- Sapp, M. (1999). *Test anxiety: Applied research, assessment, and treatment* (ed. a 2-a). Latham, MD: University Press of America.
- Spielberger, C. D., & Vagg, P. R. (Eds.). (1995). *Test anxiety: Theory, assessment, and treatment.* Washington, DC: Taylor & Francis.
- Zeidner, M. (1998). *Test anxiety: The state of the art.* New York: Plenum.

anterioară cu un numit test sau tip de teste tinde să fie avantajoasă pentru subiectul testat prin faptul că oferă un exercițiu, îi poate reduce anxietatea și îi poate crește încrederea în sine. De fapt, atunci când indivizii sunt retestați fie cu aceeași formă, fie cu o formă alternativă a unui test de aptitudine, scorurile la a doua testare sunt aproape invariabil mai mari decât primele, un fenomen cunoscut drept *efect al exersării*. Bineînțeles, dat fiind spectrul enorm al tipurilor de teste și itemi existente, un subiect poate fi foarte sofisticat și exersat în teste de un anumit tip, aplicate într-un mediu dat (ex. teste de achiziții de cunoștințe cu alegeri multiple, creion-hârtie), dar total nefamiliarizat cu teste de alt tip, aplicate în alte medii (ex. teste de aptitudini generale administrate individual sau teste de performanță administrate pe computer). Din acest motiv, subiecții testați diferă mult în privința spectrului de experiențe pe care le-au avut și care îi pot face să fie mai mult sau mai puțin pregătiți pentru o experiență de testare.

În cadrul procesului de testare, modul tradițional de a gestiona variabilitatea în sofisticare a fost acela de a oferi instrucțiuni explicite și itemi de exersare înaintea testului ca parte a procedurilor standardizate de administrare pentru a se asigura faptul că subiecții testați sunt capabili să manevreze adecvat tehnica de a răspunde la itemii testului. Deși aceste ședințe de orientare practică nu pot în niciun caz să anuleze diferențele dintre subiecții testați, ele pot asigura cel puțin faptul că examinații vor fi capabili să gestioneze competent procedeele de a răspunde la test. În plus, este bine întotdeauna ca examinatorii să pună întrebări despre experiențele anterioare ale subiecților cu tipul de test prezent sau cu tipul de test la care ei urmează să răspundă și să utilizeze această informație în interpretarea scorurilor la test. Bineînțeles, există alte căi pe care subiecții testați le pot utiliza pentru a dobândi

experiență și abilități în testare. Una dintre cele mai simple și mai eficiente metode pentru cei care urmează să fie testați pentru admiterea în școală și colegiu este de a se familiariza cu itemii și procedurile unor astfel de teste utilizând exemplele de teste furnizate de editorii acestor instrumente (vezi Info 6.2). Profesorii și consilierii pot fi de ajutor, oferind elevilor și clienților informații despre abilitățile de testare și orientări (vezi Scruggs & Mastropieri, 1992). Importanța pe care mulți oameni o dau pregătirii pentru testare este subliniată de faptul că antrenamentul pentru teste a devenit o industrie majoră ce oferă subiecților numeroase manuale, soft-uri, cursuri și servicii de pregătire adecvate diferitelor programe de testare. Dacă și în ce măsură subiecții testați sunt capabili să obțină îmbunătățiri semnificative ale scorurilor prin diferitele metode de antrenament și pregătire a testării este subiect de dezbatere. O descriere interesantă a unora dintre aspectele și descoperirile legate de acest subiect poate fi găsită în capitolul „Gaming the Tests: How Do Coaching and Cheating Affect Test Performance?" în cartea scrisă de Zwick (2002) *Fair Game? The Use of Standardized Admissions Tests in Higher Education*.

Un alt aspect al dezbaterii legate de pregătirea testării ține de distincția dintre exersarea intensivă care urmărește doar, sau în primul rând, creșterea scorurilor la test, pe de o parte, și predarea care se adresează obiectivelor mai largi ale curriculei, pe de altă parte. Acest aspect a fost subliniat de practica mult controversată a testării prin procură pe care multe state și localități o instituie din ce în ce mai mult în scopul de a face ca școlile publice să fie responsabile pentru învățarea elevilor. În acest caz, ca și în multe altele, considerațiile economice și politice creează condiții în care tocmai obiectivul testării poate fi subminat, iar responsabilitățile

diferitelor părți implicate în procesul educațional sunt deviate de la utilizatorii testelor la teste.

Problema disimulării de către subiectul testat

O perspectivă total diferită asupra predispozițiilor subiecților testați este prezentată în situațiile în care tipul de luare a deciziilor pentru care este utilizat testul promovează disimularea. Încercările subiecților testați de a se prezenta într-o lumină favorabilă sau nefavorabilă, nerealistă, nu sunt neobișnuite și pot să fie sau să nu fie conștiente. Scalele de validitate construite pentru a detecta diferite tipuri de încercări de a impresiona sau seturi de răspunsuri cum ar fi cele ce țin de defensivitate sunt încorporate în diferite inventare de personalitate (MMPI-2 și MCMI-III) și au o lungă istorie. În ultimii ani, instrumentele construite în special pentru a evalua posibilitatea de a disimula la testele cognitive administrate în contextul evaluării judiciare sau neuropsihologice — cum este Validity Indicator Profile — au fost adăugate repertoriului de instrumente disponibile în acest scop (vezi R. Rogers, 1997). Fără îndoială, obținerea cooperării complete a subiecților testați în procesul de testare este un aspect crucial de care depind fidelitatea și corectitudinea rezultatelor obținute la test (vezi Info 7.3).

Pentru o trecere în revistă a cercetărilor pe numeroase alte teme importante cu privire la perspectiva subiectului asupra testării psihologice sau educaționale, cititorii pot consulta volumul editat de Nevo și Jäger (1993) pe această temă. Studiile prezentate în această lucrare au fost realizate de cercetători din Germania, Israel și Statele Unite care au căutat să obțină feedback-uri sub forma opiniilor, atitudinilor și reacțiilor examinaților la diferite aspecte ale experienței lor de testare, având drept obiectiv final îmbunătățirea anumitor teste și a testării în general.

Obținerea consimțământului informat

Conform cu *Ethical Principles of Psychologists and Code of Conduct* (APA, 2002), înaintea administrării unui test psihologic sau a unei proceduri de evaluare, psihologii trebuie să obțină și să documenteze *consimțământul informat* al subiecților testați fie oral, fie în scris. Pentru a fi considerat *informat*, consimțământul obținut de la subiecți trebuie să fie precedat de o explicație potrivită a naturii și scopului evaluării, și de informații cu privire la limitele confidențialității și ale modului în care va fi asigurată securitatea rezultatelor testului. Alte aspecte practice care pot afecta subiecții testați într-o situație specifică (ex.: tarife, capacitatea de a refuza sau a întrerupe testarea, implicarea unei a treia părți etc.) trebuie, de asemenea, să fie discutate. Subiecților testați ar trebui să li se dea ocazia de a pune întrebări despre procesul de testare, întrebări cărora trebuie să li se răspundă. Codul etic al psihologilor (APA, 2002) enumeră câteva circumstanțe excepționale în care consimțământul informat nu este cerut fie pentru că este implicat, fie pentru că evaluarea este obligatorie. De asemenea, în acest cod sunt evidențiate recomandări suplimentare care țin de astfel de cazuri și de cazul persoanelor care nu au capacitatea de a consimți. Alți profesioniști implicați în utilizarea testării psihologice și a instrumentelor de evaluare au instrucțiuni similare pentru obținerea consimțământului informat (vezi American Counseling Association, 1995).

Importanța pregătirii examinatorului

Nu poate fi subliniat îndeajuns rolul-cheie pe care îl joacă administrarea și scorarea adecvată a testelor în acumularea datelor

care urmează să fie interpretate. Procedurile standardizate pentru administrarea și scorarea unui test specificate în manual oferă fundamentul care permite aplicarea unui cadru de referință normativ sau bazat pe criteriu în interpretarea scorurilor obținute la test. Cu siguranță, abaterile de la procedurile standardizate sau modificările acestora sunt uneori inevitabile sau necesare, așa cum se întâmplă, de exemplu, atunci când administrarea unui test este întreruptă de un anumit eveniment sau când sunt necesare anumite modificări pentru subiecții ce prezintă dizabilități. În astfel de cazuri, examinatorii trebuie să înregistreze modificările care au fost făcute. Mai mult, în măsura în care există motive să credem că scorurile obținute la un test au putut fi influențate de întreruperi sau modificări ale procedurilor standardizate, natura întreruperilor sau a modificărilor trebuie să fie transmisă celor care vor interpreta scorurile obținute la test sau vor lua decizii pe baza acestora (AERA, APA, NCME, 1999).

Cu privire la pregătirea examinatorilor, persoana care administrează un test trebuie să fie complet familiarizată cu scopul și procedurile testului, capabilă să stabilească un raport cu persoanele testate și gata să răspundă la întrebările acestora sau să se descurce în orice situație de urgență previzibilă care ar putea apărea în timpul testării. În general, administrarea testelor în grup nu necesită o instruire suplimentară celor menționate mai sus. Pe de altă parte, testarea individuală implică inevitabil o pregătire mult mai mare și, de obicei, necesită o experiență supervizată. De exemplu, atunci când întrebările testului sunt prezentate oral, ele trebuie rostite verbatim. Pentru ca administrarea să decurgă lin, să înainteze într-un ritm bun și să permită examinatorilor oportunitatea de a fi atenți și a înregistra răspunsurile (tot verbatim), examinatorii trebuie să păstreze în minte nu numai itemii și succesiunea lor, dar și un număr de reguli suplimentare care privesc tempoul prezentării și al răspunsului la itemi, punctele de început și de oprire, precum și regulile de scorare a răspunsurilor în timp ce acestea sunt date. Atunci când materiale cum ar fi piese de puzzle, cuburi, imagini și alte astfel de obiecte sunt utilizate în testarea individuală, ele trebuie prezentate subiectului și îndepărtate exact în maniera specificată în manual. Dacă examinatorii nu memorează și nu exersează toate procedurile de testare înaintea testării până când le stăpânesc complet, administrarea testului și posibilitatea subiectului de a obține o anumită performanță vor fi afectate în moduri care ar putea pune în pericol cu ușurință validitatea datelor testului.

Administrarea testelor pe computer

O cale sigură de a evita apariția erorilor în administrarea testelor și probabilitatea ca performanța la test să fie influențată de variabile care țin de sexul, rasa, vârsta, aparența, stilul interpersonal și alte variabile ale examinatorului este administrarea testelor pe baza computerului. Avantajele acestui mediu de testare cu privire la relativa uniformitate a prezentării materialelor testului și a preciziei cu care răspunsurile pot fi temporizate, înregistrate și scorate sunt evidente. De fapt, curentul înlocuirii testelor creion-hârtie cu administrarea testelor pe computer progresează, mai ales în cadrul programelor de testare pe scară largă, și este probabil să continue să se extindă pe măsură ce administrarea pe computer devine mai eficientă din punctul de vedere al costurilor. Mai mult, și dezvoltarea și utilizarea unor teste — cum ar fi Test of Variables of Attention — care prin natura lor pot exista numai în versiune pe computer progresează rapid. Pe de altă parte, administrarea testelor individuale de aptitudini generale,

tehnici proiective și multe teste neuropsihologice (dacă administrarea este realizată de un examinator capabil) are anumite avantaje bine determinate în termenii datelor calitative care pot fi acumulate, mai ales în contextul evaluărilor clinice, neuropsihologice și judiciare. Înainte de instituirea unei schimbări a metodelor tradiționale de administrare a testelor în administrarea computerizată, trebuie să fie deplin investigată comparabilitatea rezultatelor în cele două seturi de condiții — care variază în funcție de tipul de instrument (Mead & Drasgow, 1993). Pentru mai multe informații cu privire la evaluarea psihologică computerizată, inclusiv aspectele echivalenței administrării standard versus administrare computerizată, vezi Butcher (2003).

Scorarea testelor

Ne putem reaminti că diferențele inter-evaluatori și problema asociată a fidelității inter-evaluatori — discutate în Capitolul 4 printre posibilele surse ale erorii de măsurare — au fost considerate drept relevante numai pentru scorarea răspunsurilor cu final deschis, în care joacă un rol important subiectivitatea celui care scorează. Când ne aflăm în acest caz, utilizatorii testului trebuie să stabilească din punct de vedere empiric dacă cei care vor scora astfel de răspunsuri — provenite din testare individuală sau de grup — sunt suficient de bine instruiți pentru a obține rezultate aproape identice cu cele la care ar ajunge un evaluator independent și experimentat. Ca regulă generală, atunci când pot fi calculați coeficienți de fidelitate inter-evaluatori, aceștia trebuie să se apropie de +1,00 și nu trebuie să fie mai mici de 0,90. Atunci când scorurile sunt exprimate într-o altă manieră, ar trebui să existe obiectivul de a se ajunge cât mai aproape posibil la un acord de 100% între scorurile atribuite de evaluatori independenți,

instruiți. Trebuie observat că ușurința cu care astfel de standarde înalte de precizie pot fi atinse atunci când scorarea presupune judecăți subiective diferă mult între diferite tipuri de teste. Aceasta deoarece instrucțiunile pentru scorarea anumitor răspunsuri cu final deschis (răspunsuri de tip eseu la testele de achiziții de cunoștințe) tind să fie mai clare și mai ușor de stăpânit decât altele (răspunsurile la tehnici proiective, cum este testul Rorschach). Pentru o prezentare a aspectelor legate de fidelitatea evaluatorului pentru Rorschach Comprehensive System, vezi Acklin, McDowell, Verschell și Chan (2000).

În cazul testelor scorate obiectiv — pur și simplu prin numărarea răspunsurilor la diferite categorii și realizarea unor calcule cerute pentru transformarea scorurilor brute în alte forme numerice, cum ar fi scorurile standard de diferite tipuri — nu ar trebui să apară erori de scorare. O cale de a le evita este administrarea și scorarea testelor cu ajutorul computerului. Alta este utilizarea scannerelor optice și a unui program adecvat, deși este nevoie, înaintea scanării, de o examinare atentă a foilor de răspuns pentru ștersături incomplete și orice alte probleme de acest fel. Dacă scorarea obiectivă este realizată manual, mai ales atunci când sunt utilizate grile care sunt suprapuse pe foaia de răspuns, posibilitatea apariției erorilor trebuie să fie împiedicată prin sublinierea importanței absolute a preciziei și prin instituirea procedurilor de verificare dublă a tuturor calculelor necesare și a transformărilor scorurilor.

Transformarea scorurilor brute în scoruri standard la testele administrate individual cum sunt scalele Wechsler presupune, în mod tipic, o serie de calcule aritmetice și o căutare a echivalenților scorurilor brute în diferite tabele incluse în manuale. Pentru a preveni ca unele erori să intre în scoruri ca rezultat al

neglijenței în desfășurarea acestor proceduri, editurile care publică teste oferă, de asemenea, și programe care pot realiza toate calculele și transformările necesare pentru scorurile brute obținute de examinator care sunt introduse (cu atenție) în calculator. Dacă utilizatorii testului nu pot să beneficieze de acest tip de ajutor, o bună practică a testării presupune reverificarea corectitudinii tuturor calculelor și scorurilor echivalente obținute din tabele.

Aspecte esențiale ale interpretării scorurilor unui test și scrierea raportului psihologic

Cea mai mare parte din materialul anterior din această carte a urmărit să comunice cititorilor complexitatea pe care o presupune aplicarea adecvată a testelor psihologice și utilizarea datelor obținute. Interpretarea scorurilor unui test și comunicarea inferențelor realizate pe baza testării psihologice sunt punctele culminante ale utilizării testelor. Acestea sunt și componentele testării psihologice prin care subiecții testați fie obțin cele mai mari beneficii, fie suferă cel mai mare rău ce poate deriva din utilizarea testelor. Din moment ce toate diferitele principii implicate în interpretarea și raportarea rezultatelor obținute la un test nu pot fi transmise printr-o singură carte — ca să nu mai vorbim de o parte a unui capitol — această secțiune prezintă o perspectivă generală asupra acestor aspecte ale testării și referințe selectate pentru diferite teme legate de aceasta.

O perspectivă particulară asupra interpretării unui test

Scorurile obținute la un test psihologic oferă mai multe sau mai puține date cantitative fidele care *descriu* concentrat comportamentul

indivizilor ca răspuns la stimulii testului. În măsura în care testele sunt selectate, administrate și scorate cu atenție, ele oferă informații care pot fi utilizate în diferite moduri, de bază fiind pur și simplu plasarea performanței subiectului la test în categorii descriptive normative sau bazate pe criteriu oferite de cadrul de referință pe care îl utilizează testul (vezi Capitolul 3). Dacă sunt suficient de fidele și sunt interpretate în mod adecvat, datele scorurilor la test pot fi utile și în explicarea funcționării psihice a indivizilor și în luarea deciziilor cu privire la ei pe baza estimărilor oferite de scoruri cu privite la caracteristicile subiecților testați sau la comportamentul lor viitor (vezi Capitolul 5). Cu toate acestea, pentru a obține răspunsuri justificabile la întrebările complexe care sunt puse psihologilor, profesorilor și altor profesioniști din serviciile umane sau pe care și le pun ei înșiși, deseori sunt necesare date din surse multiple și trebuie realizată *întotdeauna* o judecată informată. Pentru a emite judecăți informate despre oameni este necesară aprecierea valorii și limitelor inerente ale cadrul de referință, fidelității și validității datelor scorurilor obținute la test și a datelor provenite din alte surse care ar putea fi utilizate într-un anumit caz. În plus, este necesară cunoașterea contextului și a ariilor specifice ale comportamentului uman relevant pentru aspectele respective. Mai mult, luarea deciziilor cu privire la oameni presupune în mod invariabil judecăți din partea celor care iau decizii și pune asupra lor responsabilitatea etică pentru consecințele deciziilor care sunt luate. Din păcate, în practica obișnuită, frecvent sunt ignorate aspectele contextuale, judecățile de valoare nu sunt recunoscute explicit și scorurile la test devin deseori singurul factor determinant în luarea deciziilor. Ca urmare, din motive legate de expeditivitate, cea mai mare parte a responsabilității pentru multe decizii este transferată nejustificat

de la utilizatorii testelor și de la cei care iau deciziile, la testele înseși. O modalitate de a contracara această problemă este înțelegerea și evaluarea implicațiilor diferenței dintre testarea și evaluarea psihologică, discutate în Capitolul 1. Această diferență este similară distincției dintre realizarea analizelor medicale, pe de o parte, și, pe de altă parte, integrarea rezultatelor analizelor cu istoricul medical și simptomele prezentate de un pacient pentru obținerea unui diagnostic și a unui plan de tratament (Handler & Meyer, 1998).

O perspectivă generală asupra punctelor puternice și a limitelor testării și a altor instrumente de evaluare prezentate anterior în acest capitol încearcă să transmită dezirabilitatea integrării a cât de multe surse de dovezi este posibil, ori de câte ori este pusă o întrebare semnificativă care necesită evaluarea indivizilor sau a grupurilor. Cu privire la interpretarea și utilizarea testelor, perspectiva particulară prezentată aici este următoarea:

1. Uneori, testele psihologice pot fi cele mai eficiente și obiective instrumente dinsponibile pentru a strânge date fidele și valide despre oameni.
2. Testarea psihologică poate fi deseori o componentă valoroasă a procesului de evaluare a indivizilor și a grupurilor.
3. Scorurile obținute la testele psihologice nu ar trebui să fie niciodată singura sursă de informații pe care să se bazeze deciziile care afectează viața unui individ.

Maniera particulară în care trebuie realizate interpretarea și raportarea scorurilor la test depinde de doi factori interrelaționați: (a) scopul în care a fost realizată testarea și (b) partea care a cerut testarea. Cu privire la ultima, trei posibilități distincte

determină modul în care datele testului sunt interpretate și comunicate.

- *Când psihologii utilizează testele în interesul lor* (ca instrumente în cercetare), ei pot interpreta datele grupate ale scorurilor la test și le pot comunica participanților la cercetare în orice manieră considerată a fi potrivită pentru scopurile investigației lor. Cerințele legale și standardele etice pentru cercetarea cu participanți umani guvernează practica din acest tip de testare și includ cerințe de obținere a consimțământului informat al participanților și de informare post-experiment (APA, 2002). Cu privire la rolul testării în cadrul proiectului de cercetare, aspecte cum ar fi alegerea instrumentelor, căile prin care sunt raportate scorurile și semnificația atribuită acestora trebuie să fie evaluate în lumina obiectivelor investigației. Cercetarea științifică și publicarea sunt în mod esențial întreprinderi cu autoguvernare care utilizează mecanismul revizuirii de către colegi pentru a evalua meritele substanțiale și metodologice ale unei lucrări propuse, aflată în desfășurare sau încheiată. Astfel, atunci când o cercetare care presupune utilizarea datelor la test este trimisă pentru recenzare sau este publicată, ea va fi acceptată sau respinsă, citată sau ignorată pe baza modului în care au fost utilizate datele — printre altele.
- *Atunci când psihologii utilizează testele în interesul clienților lor*, ei sunt singurii responsabili de interpretarea datelor obținute la test, integrarea lor cu informații provenite din alte surse și comunicarea rezultatelor într-o manieră adecvată și utilă clientului. Dacă aceste evaluări urmăresc diagnosticul, planificarea tratamentului, monitorizarea progreselor sau facilitarea schimbării, clientul este consumatorul ultim al informației și judecătorul final al beneficiilor aduse de aceasta. De fapt, în

multe astfel de situații, interpretarea scorurilor obținute la un test, atât timp cât este vorba despre implicații ale scorurilor, poate presupune un efort de colaborare între psiholog sau consilier și client (Fischer, 2000). Info 7.7. enumeră resurse care oferă în mod explicit ghidare pentru utilizarea testelor în contextul consilierii și evaluării clinice.

• *Când psihologii utilizează testele în interesul unei a treia părți,* cum ar fi o organizație sau un alt profesionist, ei acționează drept consultanți într-un proces de evaluare inițiat de ceilalți în scopuri proprii. Astfel, liniile de responsabilitate nu sunt la fel de clar stabilite ca în situațiile anterioare, deoarece subiectul testat care este evaluat nu este în mod necesar consumatorul ultim al datelor testului. Cu toate acestea, așa cum am menționat anterior, atât din perspectivă etică, cât și pentru a se proteja de posibile răspunderi, consultanții sunt responsabili să afle care este scopul consultației nu doar în termenii tipului de informație pe care o caută a treia parte, ci și ai modului în care informația va fi utilizată. Doar cunoscând aceste aspecte, profesioniștii din evaluarea psihologică pot determina dacă este necesară utilizarea testelor, cum vor fi raportate scorurile, ce alte instrumente pot fi necesare pentru a obține informațiile căutate și dacă sunt capabili sau dornici să participe la procesul consultației.

Interpretarea rezultatelor la test presupune o serie de inferențe făcute pe baza datelor acumulate din (a) eșantioanele de comportament actual (răspunsurile la itemii testului) (b) reunirea acestor eșantioane în unul sau mai multe scoruri, (c) dovezile disponibile cu privire la fidelitatea scorurilor obținute, (d) compararea scorurilor cu cadrele de referință normative sau bazate pe criteriu pe care testul le furnizează, (e) evaluarea

Info 7.7

Resurse despre utilizarea testelor în consiliere și în clinică

Articolele și cărțile care oferă instrucțiuni și exemple de interpretare și utilizare a testelor psihologice în domeniul practicii clinice sunt numeroase și volumul lor crește constant. Lucrările incluse în această listă scurtă prezintă numai câteva dintre numeroasele căi posibile de a aplica testarea în cadrul consilierii și în clinică. Resurse pentru aplicări specializate și pentru diferite populații sunt incluse în Info 7.8.

- Beutler, L. E., & Groth-Marnat, G. (Eds.). (2003). *Integrative assessment of adult personality* (ed. a 2-a). New York: Guilford.
- Fischer, C. T. (1994). *Individualizing psychological assessment*. Hillsdale, NJ: Erlbaum. (publicată inițial în 1985)
- Lowman, R. L. (1991). *The clinical practice of career assessment: Interests, abilities, and personality*. Washington, DC: American Psychological Association.
- Maruish, M. E. (Ed.). (2004). *The use of psychological testing for treatment planning and outcome assessment* (ed. a 3-a, Vol. 1–3). Mahwah, NJ: Erlbaum.

scorurilor în lumina calității datelor de validitate internă și externă disponibile, (f) contextul și situația specifică în care are loc testarea, (g) caracteristicile personale ale individului evaluat. Atunci când toate aceste surse de dovezi sunt privite în conjuncție una cu cealaltă și sunt adăugate informațiilor relevante colectate prin alte metode, ar trebui să fie clare implicațiile și limitele lor cu privire la întrebările specifice puse de evaluare. Ocazional (ex.: când datele din diferite surse fidele și presupus valide sunt reciproc contradictorii sau când există motive de a crede că unele dintre piesele-cheie ale dovezilor — derivate din teste sau din alte surse — nu prezintă fidelitate)

trebuie comunicată în raport natura neconcludentă a rezultatelor. În astfel de cazuri, recomandările adecvate ar putea include trimiteri către alți profesioniști sau colectarea de date suplimentare. Info 7.8 enumeră câteva dintre numeroasele resurse disponibile în prezent despre interpretarea testelor și alte subiecte legate de aceasta.

Comunicarea rezultatelor obținute la test și a rezultatelor evaluării psihologice

Indicația de bază care trebuie urmată în comunicarea rezultatelor la test este de a oferi informațiile derivate din scorurile la test, inclusiv limitele acestuia, într-un limbaj pe care receptorul îl poate înțelege. Cu toate acestea, maniera specifică în care scorurile sunt raportate poate varia larg în funcție de testele administrate, de cadrul sau contextul în care are loc testarea, de scopurile pentru care a fost realizată testarea și de cei care receptează informația. Prin urmare, modul adecvat de a raporta rezultatele testării psihologice sau rezultatele unei evaluări nu poate fi condensat într-un singur set de reguli, potrivite pentru toate cazurile. Cu toate acestea, aspecte care țin de diferite moduri de a comunica rezultatele testului sunt prezentate în paragrafele următoare.

- *Atunci când rezultatele obținute la test sunt comunicate direct subiecților testați sub forma foilor de raportare a scorurilor sau a profilelor produse de computer*, organizația responsabilă pentru programul de testare trebuie să ofere informații despre interpretarea adecvată. De exemplu, site-ul College Board Web (la http://www.collegeboard.com/student/testing/sat/scores/understanding. html) conține un material relevant

Info 7.8

Alte informații despre interpretarea testelor și evaluarea psihologică

Sursele principale de informație despre utilizarea și interpretarea unor teste și instrumente de evaluare specifice sunt manualele, ghidurile și alte documente oferite de autorii testului și editurile care l-au publicat. În plus, instrucțiuni pentru interpretarea testelor pot fi găsite în numeroase cărți și reviste. Iată câteva exemple.

Pentru o perspectivă generală asupra interpretării testelor:

- Lichtenberg, J. W., & Goodyear, R. K. (Eds.). (1999). *Scientist-practitioner perspectives on test interpretation.* Boston: Allyn & Bacon.

Pentru probleme legate de evaluarea copiilor și adolescenților:

- Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (ed. a 2-a). Boston: Allyn & Bacon.
- Kamphaus, R. W. & Frick, P. J. (2002). *Clinical assessment of child and adolescent personality and behavior* (ed. a 2-a). Boston: Allyn & Bacon.
- Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (ed. a 4-a). San Diego, CA: Author.
- Sattler, J. M. (2002). *Assessment of children: Behavioral and clinical applications* (ed. a 4-a). San Diego, CA: Author.

Pentru perspective asupra interpretării testelor în cazul unor diferite populații:

- Sandoval, J., Frisby, C. L., Geisinger, K. F., Scheuneman, J. D., & Grenier, J. R. (1998). *Test interpretation and diversity: Achieving equity in assessment.* Washington, DC: American Psychological Association.
- Ekstrom, R. B., & Smith, D. K. (Eds.). (2002). *Assessing individuals with disabilities in educational, employment, and counseling settings.* Washington, DC: American Psychological Association.

pentru interpretarea scorurilor SAT în rubrica „Cum să-ți înțelegi scorurile". Sunt disponibile suplimente similare pentru interpretarea scorurilor obținute în cadrul altor programe de testare pe scară largă, cum ar fi ACT (la http://www.act.org/aap/scores/under.html).

• *Interpretarea scorurilor obținute din programe pe computer* aplică reguli de decizie bazate pe experiența clinică și pe judecata experților, pe abordări actuariale bazate pe asocieri statistice și corelații între scoruri și criterii sau pe ambele. Modul în care sunt utilizate aceste interpretări ale testelor bazate pe computer (CBTI) și persoanele care le utilizează reprezintă tema unor discuții și dezbateri continue în cadrul profesiei de testare. Destinate inițial să ajute clinicienii la interpretarea inventarelor de personalitate și a altor instrumente diagnostice, CBTI s-au înmulțit rapid și s-au extins la diferite tipuri de teste și de utilizări ale testelor. Una dintre problemele care apar datorită acestei dezvoltări este faptul că serviciile CBTI sunt deseori disponibile comercial indivizilor care par să aibe suficiente acreditări — cum ar fi licența în practicarea psihologiei sau un grad medical — dar care nu cunosc suficient de bine testul în chestiune. În consecință, rapoartele generate de CBTI sunt mult prea des privite fără ochi critic și utilizate ca substitut al evaluării individualizate care se bazează pe surse multiple de date și este structurată de judecata unui profesionist calificat. În plus, regulile specifice utilizate în generarea unor astfel de rapoarte sunt supuse drepturilor de proprietate, ceea ce împiedică deseori o evaluare adecvată a validității interpretărilor lor de către viitorii utilizatori. Cu toate acestea, profesioniștii care beneficiază de serviciile CBTI au aceleași obligații etice cu privire la competența utilizării acestor teste

ca și în cazul altor instrumente de evaluare. Furnizorii de servicii CBTI, la rândul lor, au obligația etică de a oferi utilizatorilor informații cu privire la surse, baze de date și limite ale interpretărilor pe care le oferă (vezi APA, 2002; Moreland, 1991).

• *Mijlocul tradițional de comunicare a rezultatelor obținute la test și a rezultatelor evaluării este raportul psihologic scris.* Această abordare individualizată permite un mare grad de flexibilitate în ajustarea comunicării la scopul evaluării și la nevoile clienților și consumatorilor de date ale testelor psihologice. De asemenea, ajută evaluatorul să organizeze, să clarifice și să sintetizeze informația obținută din toate sursele disponibile și să creeze o înregistrare a rezultatelor evaluării care pot fi consultate în viitor. O descriere completă a diferitelor abordări ale scrierii raportului psihologic și a beneficiilor și capcanelor acestuia, precum și numeroase exemple despre cum să scrii și cum să nu scrii astfel de rapoarte sunt oferite de Norman Tallent (1993). Un ghid mai scurt, pas-cu-pas, de a scrie raportul psihologic pregătit de Raymond Ownby (1997) merită atenția cititorilor interesați de acest subiect. Info 7.9 enumeră câteva exemple despre ce se face și ce nu se face în interpretarea testului, deosebit de relevante pentru scrierea raportului psihologic.

Protejarea datelor testului

Păstrarea în siguranță a datelor testului, fie că acestea constau din înregistrări individuale identificabile, scoruri sau materiale ale testului (ex.: caiete, forme, întrebări, grile de scorare, manuale etc.) este responsabilitatea primară a utilizatorilor testului și a instituțiilor care controlează accesul la astfel de date. În anumite

Info 7.9

Ce se face și ce nu se face în interpretarea testelor

Câteva exemple cu privire la ceea ce *nu* este o interpretare a testului

- *Raportarea scorurilor numerice*: fie că scorurile sunt exprimate ca ranguri percentilice, IQ, scoruri în procente sau în alt format, pur și simplu enumerarea lor nu reușește să transmită semnificația și implicațiile pe care le au.
- *Atribuirea de etichete*: plasarea indivizilor în cadrul categoriilor de diagnostic (ex.: retardare mentală moderată sau personalitate de tip borderline) sau a anumitor tipologii pe baza scorurilor la test nu este un substitut adecvat pentru interpretare care amplifică înțelegerea.
- *Formularea rezultatelor în termenii unor generalități triviale:* multe afirmații se pot aplica la fel de bine tuturor indivizilor din anumite categorii (pacienți de psihiatrie, copii etc.) datorită ratelor de bază înalte din acele populații. Meehl (1956) a propus formularea „efect Barnum" pentru a caracteriza natura lipsită de valoare a descrierilor de acest tip.

Ce ar trebui să fie interpretarea unui test

- *La minim,* interpretarea scorurilor obținute la un test realizată pentru consumatorii datelor unui test ar trebui să includă o explicație clară a (a) ceea ce acoperă testul, (b) semnificației scorurilor, (c) limitelor preciziei scorurilor care derivă din eroarea de măsurare, (d) interpretării greșite obișnuite ale anumitor scoruri particulare — de exemplu IQ și (e) modului în care rezultatele la test pot sau vor fi utilizate (AERA, APA, NCME, 1999, cap. 5).
- *În cel mai bun caz,* interpretarea scorurilor la test adaugă valoare eșantioanelor de comportament obținute cu ajutorul testelor, le reunește, le integrează cu toate celelalte date disponibile și utilizează judecata informată profesional pentru a face inferențe utile și valide din punct de vedere ecologic.

cazuri, această responsabilitate poate deveni dificil de îndeplinit deoarece (a) multe cerințe legale, cerințe instituționale, standarde profesionale și preocupări etice diferite și în apariție guvernează deciziile legate de când, cum și cui se pot dezvălui datele testului și (b) aceste restricții pot diferi substanțial în funcție de cadrul și de scopul testării și pot chiar să intre în conflict una cu alta. Cel mai recent, de exemplu, Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, care a intrat în vigoare în Statele Unite pe 14 aprilie 2003, a impus cerințe care ridică multe întrebări și creează o anumită confuzie în rândul specialiștilor în testare și a altor practicieni din asistența sănătății. Anumite răspunsuri la întrebări frecvente și alte informații cu privire la modul în care HIPAA influențează psihologii pot fi găsite pe site-ul APA Insurance Trust Web la http://www.apait.org/resources/hipaa/.

Ca regulă generală, reglementările legale privind situațiile în care datele unui test pot fi făcute disponibile au prioritate față de regulamentele și interesele altor părți. Cu toate acestea, utilizatorii testelor trebuie să fie conștienți de îndatoririle lor de (a) a proteja confidențialitatea rezultatelor la test și siguranța materialelor testului în cadre și situații specifice, (b) a informa părțile implicate despre aceste îndatoriri și (c) a încerca să rezolve conflicte între diferite obligații într-o manieră concordantă cu principiile etice și cu standardele profesiei lor. Pe lângă *Testing Standards* (AERA, APA, NCME, 1999) și *Ethical Principles of Psychologists and Code of Conduct* (APA, 2002), organizațiile de psihologie au emis un număr de documente pentru a ajuta psihologii să identifice și să se confrunte cu aspecte legate de siguranța testelor și a datelor testelor și să prevină utilizarea lor greșită. Unele dintre cele mai relevante documente de acest tip sunt enumerate în Info 7.10.

Un nou mediu: testarea pe internet

Avansarea și expansiunea rapidă a internetului în ultimele două decenii a revoluționat domeniul testării psihologice la fel de mult cum a influențat aproape orice alt aspect al societății contemporane. Poate că cititorii au observat, răspândite prin această carte, un număr mare de referințe de pe site-urile internet care

Info 7.10

Informații despre păstrarea în siguranță a testelor și a datelor testării

Pe lângă *Testing Standards* (AERA, APA, NCME, 1999) și *Ethical Principles of Psychologists and Code of Conduct* (APA, 2002), care se ocupă de aspecte privind siguranța materialelor testului și a datelor testării, American Psychological Association (APA) a oferit linii directoare pentru utilizatorii de teste cu privire la aceste aspecte în alte câteva documente. Printre cele mai relevante dintre acestea se numără următoarele:

- American Psychological Association, Committee on Legal Issues. (1996). Strategies for private practitioners coping with subpoenas or compelled testimony for client records or test data. *Professional Psychology: Research and Practice, 27,* 245–251.
- American Psychological Association, Committee on Psychological Tests and Assessment. (1996). Statement on the disclosure of test data. *American Psychologist, 51,* 644–648.
- American Psychological Association, Committee on Psychological Tests and Assessment. (2003). *Statement on the use of secure psychological tests in the education of graduate and undergraduate psychology students.* Accesată pe 19 februarie, 2003, de la http://www.apa.org/science/securetests.html.
- Test security: Protecting the integrity of tests. (1999). *American Psychologist, 54,* 1078.

oferă informații despre testele psihologice și aspectele legate de testare. Pe lângă acestea, și multe alte surse de informație despre teste și testare disponibile în mod curent online, livrarea serviciilor de testare psihologică pe internet de către diferiți furnizori — unii legitimi, alții nu la fel de mult — este în dezvoltare. Însă impactul pe care internetul este probabil să-l aibă asupra testării promite să fie mult mai mare în deceniile care urmează în primul rând datorită vitezei cu care testele pot fi construite, normate, publicate și revizuite utilizând capacitățile World Wide Web, cât și datorită eficienței și costurilor reduse cu care serviciile de testare pot fi livrate online.

În orice aspect al utilizării internetului, fie pentru comerț, distracție, informare sau comunicare, natura puternic democratică și accesibilă a mediului aduce atât avantaje, cât și pericole. Așa cum cititorul a putut înțelege, posibilitățile internetului cu privire la testare se orientează atât în direcția utilizărilor vătămătoare, cât și a potențialului pentru progrese excepționale. Ca urmare, a început deja examinarea profesionistă a dificultăților și a beneficiilor inerente în practica testării psihologice prezentă și previzibilă pe internet (vezi Buchanan, 2002). Unul dintre cele mai ample eforturi în această privință până în acest moment este raportul pregătit de APA Task Force on Psychological Testing on the Internet (Naglieri et al., 2004). Acest raport se concentrează asupra diferitelor aspecte legate de influența internetului asupra testării și evaluării psihologice, inclusiv asupra aspectelor psihometrice, etice și legale și a considerațiilor practice cu privire la siguranța testelor și accesul la teste. Conține, de asemenea, exemple care ilustrează aplicațiile curente ale testării pe internet și oferă indicii ale posibilităților viitoare permise de acest mediu și recomandări pentru viitor.

Aplicație practică

Cum să vă testați capacitatea de a evalua testele psihologice

Diferite site-uri web oferă anchete, chestionare și alte instrumente despre care susțin că sunt teste psihologice. Aproape orice motor de căutare vă va conduce la aceste site-uri atunci când introduceți termenii „psychological tests" sau „personality tests." Printre numeroasele oferte ale acestor site-uri se află o largă varietate de instrumente, cum ar fi „teste IQ", „teste EQ" și „teste de carieră" care pot arăta ca testele psihologice și sunt atrăgătoare pentru nespecialiști, dar au o bază științifică redusă sau total absentă. Unele dintre aceste așa-numite teste sunt oferite gratuit; altele presupun plata unei taxe modeste (de obicei) de 5, 10 sau 15 dolari. În numeroase cazuri, instrumentele gratuite nu includ rezultatele complete (sau chiar niciun rezultat) dacă nu este plătită această taxă. Deseori subiecții testați nu sunt informați cu privire la taxele necesare decât după ce au completat testul. Dacă ați citit cu atenție acest capitol, ar trebui să fiți deja conștienți de faptul că cei care publică teste psihologice legitime sunt extrem de preocupați de aspectele legate de siguranța testelor și nu doresc să-și vadă produsele răspândite pe scară largă sau utilizate inadecvat. Cu rare excepții (ex.: instrumente care sunt în proces de dezvoltare) este improbabil să găsești editori serioși sau autori care să-și facă disponibile produsele pe internet, gratuit.

O modalitate de a aplica unele dintre cunoștințele pe care le-ați acumulat citind această carte este de a inspecta unele dintre instrumentele oferite pe internet și de a le evalua în următoarele privințe (a) dacă îndeplinesc criteriile care le-ar putea califica drept teste psihologice legitime și (b) dacă informațiile pe care le oferă au valoare.

Concluzii

Să luăm în considerare următoarele întrebări și probleme tipice:

- Copilul meu are tulburare de deficit de atenție?
- Cum ar trebui să procedăm pentru a selecta cei mai buni candidați pentru departamentul de poliție?
- Acest pacient în vârstă de 65 de ani suferă de demență incipientă sau este deprimat datorită decesului soției sale?
- Care dintre acești trei candidați la poziția de director financiar, cu experiență la fel de mare în companie, ar trebui angajat?
- În ce domeniu ar trebui să mă specializez?
- Ar trebui eliberat din închisoare acest deținut condamnat pentru abuz sexual?
- Ce tip de intervenție terapeutică ar fi cea mai eficientă pentru acest client?
- Acest individ pretinde că are o dizabilitate pentru a obține compensații sau prezintă într-adevăr o dizabilitate?

Deși toate aceste întrebări și probleme de evaluare pot fi abordate cu ajutorul testelor, niciuneia nu i se poate răspunde printr-un singur scor obținut la un test sau chiar printr-o combinație de scoruri. În aceste situații și în altele similare sunt necesare date provenite din surse multiple și informații despre obiectivele specifice ale persoanelor implicate, despre contextele în care au apărut întrebările și consecințele potențiale ale deciziilor în fiecare caz. Ar trebui să fie clar în acest punct că aplicarea testelor psihologice este amenințată la fiecare pas de posibilitatea erorii. Totuși, așa se întâmplă în cele mai multe acțiuni umane. Din

momentul în care întrebările esențiale (i.e. dacă utilizarea testelor poate contribui la luarea deciziilor într-un mod mai rațional, echitabil, benefic și responsabil într-o situație dată) este nevoie de implementarea unui plan pentru utilizarea testelor cu o conștiinciozitate cât mai mare în toate fazele sale. Dacă acest fapt este realizat, este probabil ca testele, mai ales în comparație cu alte instrumente, să se dovedească a fi utile.

Testează-te singur

1. **Prima întrebare care trebuie luată în considerare atunci când se are în vedere testarea psihologică este**
   a) dacă testarea este necesară.
   b) ce tip de test va fi folosit.
   c) care sunt costurile asociate testării.
2. **Care dintre următoarele aspecte *nu* este un avantaj principal care ține de utilizarea testelor psihologice?**
   a) Obiectivitatea
   b) Eficiența
   c) Disponibile cu ușurință
3. **Câștigul potențial în precizia unei decizii de selecție este cel mai mare atunci când rata de bază este mai aproape de**
   (a) 1,00.
   (b) 0,75.
   (c) 0,50.
   (d) 0,00.
4. **Pentru precizie în selectarea angajaților, o situație ideală va presupune o rată de bază ____, o rată de selecție _____ și un test cu validitate _____.**
   (a) moderată/înaltă/înaltă
   (b) moderată/scăzută/înaltă
   (c) înaltă/scăzută/moderată
   (d) înaltă/înaltă/înaltă
5. **Care dintre următoarele instrumente de evaluare este cel mai probabil să ofere informații de încredere pentru evaluarea individuală?**
   (a) Interviul
   (b) Recomandările
   (c) Observația informală
   (d) Datele biografice

6. **Ca regulă generală, prezența în timpul administrării testelor individuale a unei a treia părți, alta decât examinatorul și subiectul testat, este**
   (a) dezirabilă.
   (b) indezirabilă.
   (c) nici a, nici b.

7. **Care dintre următoarele aspecte *nu* reprezintă una dintre ariile în care administrarea testelor bazată pe computer oferă avantaje față de administrarea individuală a testelor?**
   (a) Uniformitatea procedurii
   (b) Eficiența costurilor
   (c) Abilități de precizie
   (d) Acumulări de date calitative

8. **Perspectiva asupra testării prezentată de autor subliniază faptul că testele psihologice**
   (a) pot fi o componentă valoroasă în cele mai multe cazuri în care este necesară evaluarea psihologică.
   (b) poate fi deseori singura sursă de informație pe care se bazează deciziile.
   (c) constituie invariabil cel mai eficient instrument disponibil pentru evaluare.

9. **În comunicarea rezultatelor la test consumatorilor datelor testului, cea mai relevantă informație care li se transmite este**
   (a) scorul numeric obținut de examinați.
   (b) etichete sau diagnostice derivate din scorurile la test.
   (c) semnificația scorurilor obținute la test.

10. **Responsabilitățile legale și etice ale utilizatorilor testului cu privire la interpretarea adecvată a rezultatelor la test sunt**
    (a) înlăturate atunci când utilizează testele în interes propriu.
    (b) înlăturate atunci când utilizează testele la cererea unei a treia părți.
    (c) înlăturate atunci când utilizează testele la cererea clienților.
    (d) niciodată înlăturate.

Răspunsuri: 1. a; 2. c; 3. c; 4. b; 5. d; 6. b; 7. d; 8. a; 9. c; 10. d

Anexa A
Testele disponibile comercial menționate în text

Numele testului (Abrevierea versiunii curente)	Codul editurii
ACT Assessment	ACT
Beck Depression Inventory (BDI)	TPC
Beta III	TPC
Boston Diagnostic Aphasia Examination (BDAS)	LWW
Bracken Basic Concept Scale–Revised (BBCS-R)	TPC
California Psychological Inventory (CPI)	CPP
Clerical Abilities Battery (CAB)	TPC
College-Level Examination Program (CLEP)	TCB
Crawford Small Parts Dexterity Test (CSPDT)	TPC
Das-Naglieri Cognitive Assessment System (CAS)	RIV
Differential Ability Scales (DAS)	TPC
Graduate Record Exam (GRE)	ETS
Halstead-Reitan Neuropsychological Battery (HRNB)	RNL
Infant-Toddler Developmental Assessment (IDA)	RIV
Iowa Test of Basic Skills (ITBS)	RIV
Jackson Vocational Interest Survey (JVIS)	SIG
Kaufman Assessment Battery for Children (K-ABC-II)	AGS

Anexa B conține numele complete și adresele de internet ale editurilor care publică testele enumerate mai sus, în ordine alfabetică, conform codului editurii utilizat în această Anexă.

Kaufman Adolescent and Adult Intelligence Test (KAIT)	AGS
Law School Admission Test (LSAT)	LSAC/LSAS
Medical College Admission Test (MCAT)	AAMC
Millon Clinical Multiaxial Inventory (MCMI-III)	PA
Mini-Mental State Examination (MMSE)	PAR
Minnesota Multiphasic Personality Inventory (MMPI-2 & MMPI-A)	UMP/PA
Myers-Briggs Type Indicator (MBTI)	CPP
Otis-Lennon School Ability Test (OLSAT 8)	HEM
Quality of Life Inventory (QOLI)	PA
Revised NEO Personality Inventory (NEO PI-R)	PAR
Rorschach	H & H
SAT (formerly known as Scholastic Aptitude Test)	TCB
Stanford-Binet Intelligence Scale (S-B 5)	RIV
Stanford Diagnostic Mathematics Test (SDMT)	HEM
Stanford Diagnostic Reading Test (SDRT)	HEM
State-Trait Anxiety Inventory (STAI)	Mind
Strong Interest Inventory (SII)	CPP
Symptom Checklist-90–Revised (SCL-90-R)	PA
Test of English as a Foreign Language (TOEFL)	ETS
Test of Variables of Attention (TOVA)	UAD
Thematic Apperception Test (TAT)	HAR
Validity Indicator Profile (VIP)	PA
Wechsler Adult Intelligence Scale (WAIS-III)	TPC
Wechsler Intelligence Scale for Children (WISC-IV)	TPC
Whitaker Index of Schizophrenic Thinking (WIST)	WPS
Wide Range Achievement Test (WRAT)	WRI
Wonderlic Personnel Test	WON
Woodcock-Johnson batteries	RIV
WorkKeys Assessments	ACT

Anexa B
Adresele editurilor care publică teste și ale distribuitorilor

Cod	Editura	Adresa
AAMC	Association of American Medical Colleges	www.aamc.org
ACT	ACT, Inc.	www.act.org
AGS	American Guidance Service	www.agsnet.com
CPP	Consulting Psychologists Press	www.cpp.com
ETS	Educational Testing Service	www.ets.org
HAR	Harvard University Press	www.hup.harvard.edu
HEM	Harcourt Educational Measurement	www.HEMWEB.com
H & H	Hogrefe & Huber Publishers	www.hhpub.com
LSAC/ LSAS	Law School Admission Council/Law School Admission Service	www.lsac.org
LWW	Lippincott Williams & Wilkins	www.lww.com
Mind	Mind Garden, Inc.	www.mindgarden.com
PA	Pearson Assessments (fost NCS)	www.pearsonassessments.com

Vezi Anexa A pentru codul editurii

PAR	Psychological Assessment Resources	www.parinc.com
RNL	Reitan Neuropsychological Laboratory	www.reitanlabs.com
RIV	Riverside Publishing	www.riversidepublishing.com
SIG	Sigma Assessment Systems, Inc.	www.sigmaassessment-systems.com
TCB	The College Board	www.collegeboard.com
TPC	The Psychological Corporation	www.PsychCorp.Com
UAD	Universal Attention Disorders, Inc.	www.tovatest.com
UMP	University of Minnesota Press	www.upress.umn.edu/tests
WON	Wonderlic, Inc.	www.wonderlic.com
WPS	Western Psychological Services	www.wpspublish.com
WRI	Wide Range, Inc.	www.widerange.com

Anexa C
Tabel al ariilor situate sub curba normală

Explicarea tabelului

Coloana (1) prezintă scorurile standard (i.e. scoruri *z*) de la 0,00 la 3,24 (la intervale de 0,01) și de la 3,30 la 3,70 (la intervale de 0,10).

$$\text{scorul } z = \frac{X}{\sigma} \qquad \text{(C.1)}$$

unde

x = distanța dintre orice punct al abscisei și *media* distribuției

σ (sigma) = *abaterea standard* a distribuției

Media scorurilor *z* este zero; abaterea standard a scorurilor *z* este 1.

Coloana (2) prezintă proporțiile din aria de sub curba normală cuprinse în segmentul dintre medie și orice scor *z*. Deoarece curba normală este perfect simetrică, dacă $z = 0$ (la medie), o jumătate din curbă (0,5000 sau 50%) se află peste *z* și o jumătate sub *z*. Atunci când curba este împărțită în oricare alt punct decât media, va exista o arie *mai mare*, prezentată în Coloana (3) și o arie *mai mică* prezentată în Coloana (4). Dacă punctul care împarte curba se află în *stânga* mediei, scorul *z* are un semn *negativ* și aria mai mică se află în stânga sa. Dacă punctul care împarte curba se află în partea *dreaptă* a mediei, scorul *z* este *pozitiv* și aria mai mică se află în dreapta sa. Coloana (5) prezintă valorile pe ordonata *y*, sau înălțimea curbei, pentru fiecare scor *z*.

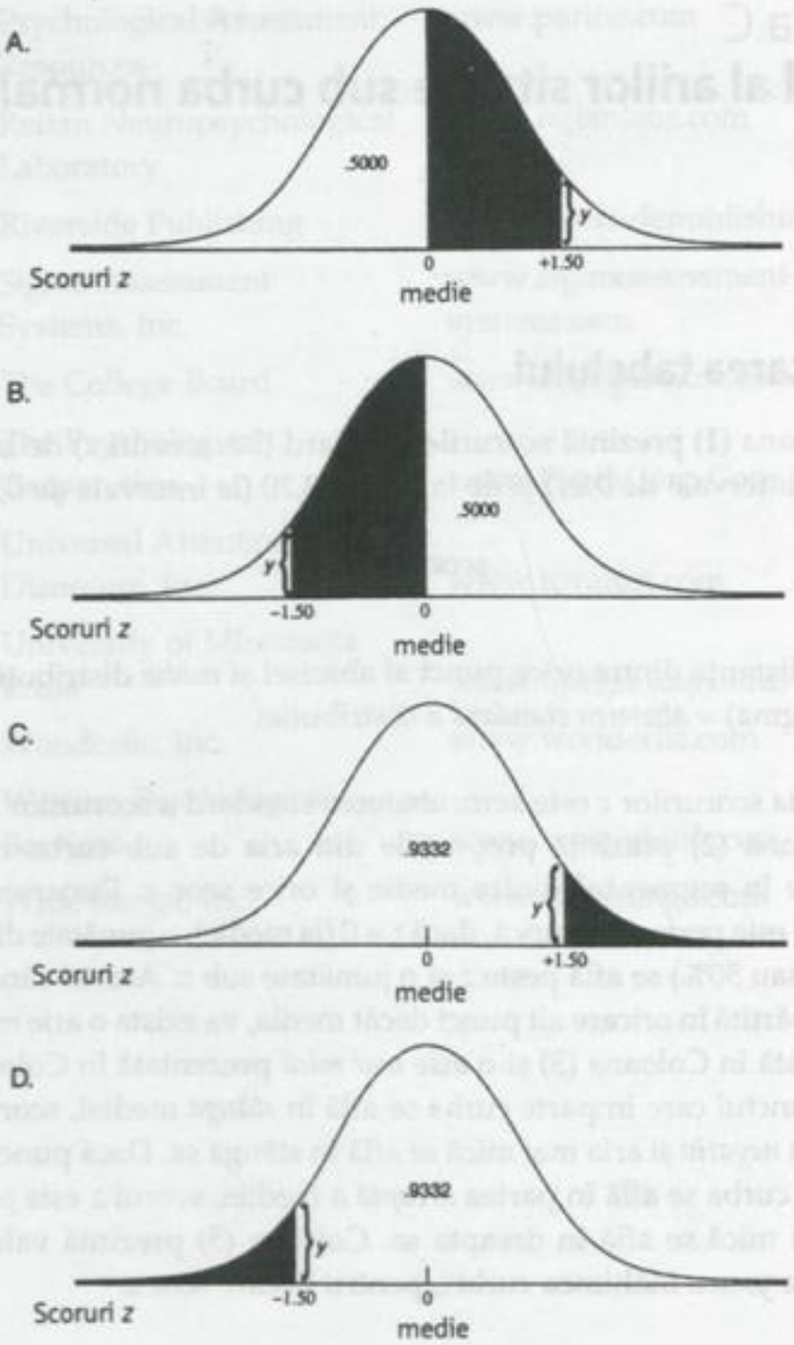

Figura C.1. Ariile situate sub curba normală

Cum se utilizează tabelul

Relațiile dintre scorurile *z* și ariile situate sub curba normală

Figura C.1, Graficul A, ilustrează distanța dintre un scor *z* de 1,50 și medie. Dacă scorul *z* de 1,50 este găsit în Coloana (1) a tabelului, Coloana (2) arată că aria dintre acest scor și medie este 0,4332 sau 43,32% din aria de sub curbă. Deoarece curba este simetrică, orice scor *z* dat subîntinde aceeași arie de la medie, fie că este pozitiv (peste medie), fie că este negativ (sub medie). Prin urmare, în Figura C.1, Graficul B, un scor *z* de -1,50 subîntinde din nou o arie de 0,4332 între medie și *z*. Pentru a afla aria aflată *sub* un scor *z* de +1,50, vom scădea 0,4332 din 0,5000 și vom obține 0,0668 sau 6,68%. Pentru a afla aria aflată *sub* un scor *z* de +1,50 vom aduna 0,4332 cu 0,5000 și vom obține 0,9332 sau 93,32%. Aceste valori sunt prezentate în Graficul C al Figurii C.1. Graficul D al figurii arată rezultatele pentru un scor *z* de -1,50. Aici, ca și în cazul tuturor scorurilor *z* negative ale unei curbe normale, porțiunea mai mare se află deasupra lui *z* și astfel poate fi găsită adunând la valoarea din Coloana (2) 0,5000 (0,4332+0,5000 = 0,9332); porțiunea mai mică se află sub *z* și poate fi găsită scăzând 0,4332 din 0,5000, rezultând 0,0668. Coloanele (3) și (4) ale tabelului oferă rezultatele acestor calcule. Pentru a verifica rezultatele, găsiți intrarea pentru un scor *z* de 1,50 în coloanele (3) și (4). Pentru a exersa în continuare, coroborați ariile de sub curbă prezentate în Figura 2.2 a Capitolului 2 pentru valorile σ sau scorurile *z* de ±1, ±2 și ±3.

Utilizarea tabelului pentru testarea ipotezelor

Atunci când curba normală este folosită pentru testarea ipotezelor, în statisticile inferențiale, ea este utilizată pentru a stabili probabilitatea ca valoarea critică (*z*) obținută să poată să fi rezultat întâmplător.

Deoarece valorile z dau proporţia din arie numai pentru un capăt al curbei, când ipotezele permit posibilitatea variaţiei în două direcţii, proporţia ariei care cade sub valorile critice z trebuie să fie dublată pentru a afla nivelul de probabilitate asociat cu valoarea critică obţinută.

Exemplu: Să presupunem că încercăm să aflăm dacă o diferenţă de 10 puncte care a fost obţinută între două scoruri IQ ale unui individ (un IQ Verbal de 115 şi IQ de performanţă de 105) este semnificativă din punct de vedere statistic în situaţia în care nu este de aşteptat ca un coeficient IQ să fie mai mare decât altul. Pentru a testa ipoteza de nul conform căreia nu există nicio diferenţă, obţinem o valoare critică z pentru diferenţa obţinută de 10 puncte prin împărţirea acestei diferenţe la eroarea standard a diferenţei SE_{dif} dintre scoruri, care este derivată statistic din coeficienţii respectivi de fidelitate ai scalelor Verbal şi Performanţă (vezi Capitolul 4). În acest exemplu, să presupunem că $SE_{dif} = 5$; raportul critic z este $10 \div 5 = 2{,}00$. Aria de sub această valoare critică z, 0,0228, reprezintă nivelul de probabilitate (*p*) pentru diferenţa obţinută. Cu toate acestea, deoarece ar fi putut să apară diferenţa scorurilor în oricare dintre cele două direcţii, *p* este dublat ($0{,}0228 \times 2 = 0{,}0456$) pentru a obţine probabilitatea (4,56%) ca o diferenţă de 10 puncte între scoruri să fi putut fi obţinută dacă nu ar fi existat nicio diferenţă între cele două scoruri IQ.

Acest exemplu descrie un *test bilateral*, care constituie modul tipic prin care este testată semnificaţia diferitelor rezultate. Dacă există o ipoteză direcţională specifică, cum ar putea fi în cazul unui experiment în care se aşteaptă ca rezultatele să se situeze într-o anumită direcţie, este realizat un test *unilateral*. În astfel de cazuri, dacă rezultatele se situează într-adevăr în direcţia aşteptată, nuvelul *p* pentru valoarea critică nu trebuie dublat.

Tabelul C.1 Tabel al ariilor situate sub curba normală şi al ordonatelor curbei normale în termenii scorurilor standard $z = x/\sigma_x$.

Scor standard $z = x/\sigma_x$ (1)	Aria de la medie la z (2)	Aria porţiunii mai mari (3)	Aria porţiunii mai mici (4)	Y pentru x/σ_x (5)
0,00	0,0000	0,5000	0,5000	0,3989
0,01	0,0040	0,5040	0,4960	0,3989
0,02	0,0080	0,5080	0,4920	0,3989
0,03	0,0120	0,5120	0,4880	0,3988
0,04	0,0160	0,5160	0,4840	0,3986
0,05	0,0199	0,5199	0,4801	0,3984
0,06	0,039	0,5239	0,4761	0,3982
0,07	0,0279	0,5279	0,4721	0,3980
0,08	0,0319	0,5319	0,4681	0,3977
0,09	0,0359	0,5359	0,4641	0,3973
0,10	0,0398	0,5398	0,4602	0,3970
0,11	0,0438	0,5438	0,4562	0,3965
0,12	0,0478	0,5478	0,4522	0,3961
0,13	0,0517	0,5517	0,4483	0,3956
0,14	0,0557	0,5557	0,4443	0,3951
0,15	0,0596	0,5596	0,4404	0,3945
0,16	0,0636	0,5636	0,4364	0,3939
0,17	0,0675	0,5675	0,4325	0,3932
0,18	0,0714	0,5714	0,4286	0,3925
0,19	0,0753	0,5753	0,4247	0,3918
0,20	0,0793	0,5793	0,4207	0,3910
0,21	0,032	0,5832	0,4168	0,3902
0,22	0,0871	0,5871	0,4129	0,3894
0,23	0,0910	0,5910	0,4090	0,3885
0,24	0,0948	0,5948	0,4052	0,3876
0,25	0,0987	0,5987	0,4013	0,3867
0,26	0,1026	0,6026	0,3974	0,3857

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porţiunii mai mari (3)	Aria porţiunii mai mici (4)	Y pentru x/σ_x (5)
0,27	0,1064	0,6064	0,3936	0,3847
0,28	0,1103	0,6103	0,3897	0,3836
0,29	0,1141	0,6141	0,3859	0,3825
0,30	0,1179	0,6179	0,3821	0,3814
0,31	0,1217	0,6217	0,3783	0,3802
0,32	0,1255	0,6255	0,3745	0,3790
0,33	0,1293	0,6293	0,3707	0,3778
0,34	0,1331	0,6331	0,3669	0,3765
0,35	0,1368	0,6368	0,3632	0,3752
0,36	0,106	0,6406	0,3594	0,3739
0,37	0,1443	0,6443	0,3557	0,3725
0,38	0,1480	0,6480	0,3520	0,3712
0,39	0,1517	0,6517	0,3483	0,3697
0,40	0,1554	0,6554	0,3446	0,3683
0,41	0,1591	0,6591	0,3409	0,3668
0,42	0,1628	0,6628	0,3372	0,3653
0,43	0,1664	0,6664	0,3336	0,3637
0,44	0,1700	0,6700	0,3300	0,3621
0,45	0,1736	0,6736	0,3264	0,3605
0,46	0,1772	0,6772	0,3228	0,3589
0,47	0,1808	0,6808	0,3192	0,3572
0,48	0,1844	0,6844	0,3156	0,3555
0,49	0,1879	0,6879	0,3121	0,3538
0,50	0,1915	0,6915	0,3085	0,3521
0,51	0,1950	0,6950	0,3050	0,3503
0,52	0,1985	0,6985	0,3015	0,3485
0,53	0,2019	0,7019	0,2981	0,3467
0,54	0,2054	0,7054	0,2946	0,3448
0,55	0,2088	0,7088	0,2912	0,3429
0,56	0,2123	0,7123	0,2877	0,3410
0,57	0,2157	0,7157	0,2843	0,3391

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porţiunii mai mari (3)	Aria porţiunii mai mici (4)	Y pentru x/σ_x (5)
0,58	0,2190	0,7190	0,2810	0,3372
0,59	0,2224	0,7224	0,2776	0,3352
0,60	0,2257	0,7257	0,2743	0,3332
0,61	0,2291	0,7291	0,2709	0,3312
0,62	0,2324	0,7324	0,2676	0,3292
0,63	0,2357	0,7357	0,2643	0,3271
0,64	0,2389	0,7389	0,2611	0,3251
0,65	0,2422	0,7422	0,2578	0,3230
0,66	0,2454	0,7454	0,2546	0,3209
0,67	0,2486	0,7486	0,2514	0,3187
0,68	0,2517	0,7517	0,2483	0,3166
0,69	0,2549	0,7549	0,2451	0,3144
0,70	0,2580	0,7580	0,2420	0,3123
0,71	0,2611	0,7611	0,2389	0,3101
0,72	0,2642	0,7642	0,2358	0,3079
0,73	0,2673	0,7673	0,2327	0,3056
0,74	0,2704	0,7704	0,2296	0,3034
0,75	0,2734	0,7734	0,2266	0,3011
0,76	0,2764	0,7764	0,2236	0,2989
0,77	0,2794	0,7794	0,2206	0,2966
0,78	0,2823	0,7823	0,2177	0,2943
0,79	0,2852	0,7852	0,2148	0,2920
0,80	0,2881	0,7881	0,2119	0,2897
0,81	0,2910	0,7910	0,2090	0,2874
0,82	0,2939	0,7939	0,2061	0,2850
0,83	0,2967	0,7967	0,2033	0,2827
0,84	0,2995	0,7995	0,2005	0,2803
0,85	0,3023	0,8023	0,1977	0,2780
0,86	0,3051	0,8051	0,1949	0,2756
0,87	0,3078	0,8078	0,1922	0,2732
0,88	0,3106	0,8106	0,1894	0,2709

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porţiunii mai mari (3)	Aria porţiunii mai mici (4)	Y pentru x/σ_x (5)
0,89	0,3133	0,8133	0,1867	0,2685
0,90	0,3159	0,8159	0,1841	0,2661
0,91	0,3186	0,8186	0,1814	0,2637
0,92	0,3212	0,8212	0,1788	0,2613
0,93	0,3238	0,8238	0,1762	0,2589
0,94	0,3264	0,8264	0,1736	0,2565
0,95	0,3289	0,8289	0,1711	0,2541
0,96	0,3315	0,8315	0,1685	0,2516
0,97	0,3340	0,8340	0,1660	0,2492
0,98	0,3365	0,8365	0,1635	0,2468
0,99	0,3389	0,8389	0,1611	0,2444
1,00	0,3413	0,8413	0,1587	0,2420
1,01	0,3438	0,8438	0,1562	0,2396
1,02	0,3461	0,8461	0,1539	0,2371
1,03	0,3485	0,8485	0,1515	0,2347
1,04	0,3508	0,8508	0,1492	0,2323
1,05	0,3531	0,8531	0,1469	0,2299
1,06	0,3554	0,8554	0,1446	0,2275
1,07	0,3577	0,8577	0,1423	0,2251
1,08	0,3599	0,8599	0,1401	0,2227
1,09	0,3621	0,8621	0,1379	0,2203
1,10	0,3643	0,8643	0,1357	0,2179
1,11	0,3665	0,8665	0,1335	0,2155
1,12	0,3686	0,8686	0,1314	0,2131
1,13	0,3708	0,8708	0,1292	0,2107
1,14	0,3729	0,8729	0,1271	0,2083
1,15	0,3749	0,8749	0,1251	0,2059
1,16	0,3770	0,8770	0,1230	0,2036
1,17	0,3790	0,8790	0,1210	0,2012
1,18	0,3810	0,8810	0,1190	0,1989
1,19	0,3830	0,8830	0,1170	0,1965

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porţiunii mai mari (3)	Aria porţiunii mai mici (4)	Y pentru x/σ_x (5)
1,20	0,3849	0,8849	0,1151	0,1942
1,21	0,3869	0,8869	0,1131	0,1919
1,22	0,3888	0,8888	0,1112	0,1895
1,23	0,3907	0,8907	0,1093	0,1872
1,24	0,3925	0,8925	0,1075	0,1849
1,25	0,3944	0,8944	0,1056	0,1826
1,26	0,3962	0,8962	0,1038	0,1804
1,27	0,3980	0,8980	0,1020	0,1781
1,28	0,3997	0,8997	0,1003	0,1758
1,29	0,4015	0,9015	0,0985	0,1736
1,30	0,4032	0,9032	0,0968	0,1714
1,31	0,4049	0,9049	0,0951	0,1691
1,32	0,4066	0,9066	0,0934	0,1669
1,33	0,4082	0,9082	0,0918	0,1647
1,34	0,4099	0,9099	0,0901	0,1626
1,35	0,4115	0,9115	0,0885	0,1604
1,36	0,4131	0,9131	0,0869	0,1582
1,37	0,4147	0,9147	0,0853	0,1561
1,38	0,4162	0,9162	0,0838	0,1539
1,39	0,4177	0,9177	0,0823	0,1518
1,40	0,4192	0,9192	0,0808	0,1497
1,41	0,4207	0,9207	0,0793	0,1476
1,42	0,4222	0,9222	0,0778	0,1456
1,43	0,4236	0,9236	0,0764	0,1435
1,44	0,4251	0,9251	0,0749	0,1415
1,45	0,4265	0,9265	0,0735	0,1394
1,46	0,4279	0,9279	0,0721	0,1374
1,47	0,4292	0,9292	0,0708	0,1354
1,48	0,4306	0,9306	0,0694	0,1334
1,49	0,4319	0,9319	0,0681	0,1315
1,50	0,4332	0,9332	0,0668	0,1295

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porțiunii mai mari (3)	Aria porțiunii mai mici (4)	Y pentru x/σ_x (5)
1,51	0,4345	0,9345	0,0655	0,1276
1,52	0,4357	0,9357	0,0643	0,1257
1,53	0,4370	0,9370	0,0630	0,1238
1,54	0,4382	0,9382	0,0618	0,1219
1,55	0,4394	0,9394	0,0606	0,1200
1,56	0,4406	0,9406	0,0594	0,1182
1,57	0,4418	0,9418	0,0582	0,1163
1,58	0,4429	0,9429	0,0571	0,1145
1,59	0,4441	0,9441	0,0559	0,1127
1,60	0,4452	0,9452	0,0548	0,1109
1,61	0,4463	0,9463	0,0537	0,1092
1,62	0,4474	0,9474	0,0526	0,1074
1,63	0,4484	0,9484	0,0516	0,1057
1,64	0,4495	0,9495	0,0505	0,1040
1,65	0,4505	0,9505	0,0495	0,1023
1,66	0,4515	0,9515	0,0485	0,1006
1,67	0,4525	0,9525	0,0475	0,0989
1,68	0,4535	0,9535	0,0465	0,0973
1,69	0,4545	0,9545	0,0455	0,0957
1,70	0,4554	0,9554	0,0446	0,0940
1,71	0,4564	0,9564	0,0436	0,0925
1,72	0,4573	0,9573	0,0427	0,0909
1,73	0,4582	0,9582	0,0418	0,0893
1,74	0,4591	0,9591	0,0409	0,0878
1,75	0,4599	0,9599	0,0401	0,0863
1,76	0,4608	0,9608	0,0392	0,0848
1,77	0,4616	0,9616	0,0384	0,0833
1,78	0,4625	0,9625	0,0375	0,0818
1,79	0,4633	0,9633	0,0367	0,0804
1,80	0,4641	0,9641	0,0359	0,0790
1,81	0,4649	0,9649	0,0351	0,0775

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porțiunii mai mari (3)	Aria porțiunii mai mici (4)	Y pentru x/σ_x (5)
1,82	0,4656	0,9656	0,0344	0,0761
1,83	0,4664	0,9664	0,0336	0,0748
1,84	0,4671	0,9671	0,0329	0,0734
1,85	0,4678	0,9678	0,0322	0,0721
1,86	0,4686	0,9686	0,0314	0,0707
1,87	0,4693	0,9693	0,0307	0,0694
1,88	0,4699	0,9699	0,0301	0,0681
1,89	0,4706	0,9706	0,0294	0,0669
1,90	0,4713	0,9713	0,0287	0,0656
1,91	0,4719	0,9719	0,0281	0,0644
1,92	0,4726	0,9726	0,0274	0,0632
1,93	0,4732	0,9732	0,0268	0,0620
1,94	0,4738	0,9738	0,0262	0,0608
1,95	0,4744	0,9744	0,0256	0,0596
1,96	0,4750	0,9750	0,0250	0,0584
1,97	0,4756	0,9756	0,0244	0,0573
1,98	0,4761	0,9761	0,0239	0,0562
1,99	0,4767	0,9767	0,0233	0,0551
2,00	0,4772	0,9772	0,0228	0,0540
2,01	0,4778	0,9778	0,0222	0,0529
2,02	0,4783	0,9783	0,0217	0,0519
2,03	0,4788	0,9788	0,0212	0,0508
2,04	0,4793	0,9793	0,0207	0,0498
2,05	0,4798	0,9798	0,0202	0,0488
2,06	0,4803	0,9803	0,0197	0,0478
2,07	0,4808	0,9808	0,0192	0,0468
2,08	0,4812	0,9812	0,0188	0,0459
2,09	0,4817	0,9817	0,0183	0,0449
2,10	0,4821	0,9821	0,0179	0,0440
2,11	0,4826	0,9826	0,0174	0,0431
2,12	0,4830	0,9830	0,0170	0,0422

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porțiunii mai mari (3)	Aria porțiunii mai mici (4)	*Y* pentru x/σ_x *(5)*
2,13	0,4834	0,9834	0,0166	0,0413
2,14	0,4838	0,9838	0,0162	0,0404
2,15	0,4842	0,9842	0,0158	0,0396
2,16	0,446	0,9846	0,0154	0,0387
2,17	0,4850	0,9850	0,0150	0,0379
2,18	0,4854	0,9854	0,0146	0,0371
2,19	0,4857	0,9857	0,0143	0,0363
2,20	0,4861	0,9861	0,0139	0,0355
2,21	0,4864	0,9864	0,0136	0,0347
2,22	0,4868	0,9868	0,0132	0,0339
2,23	0,4871	0,9871	0,0129	0,0332
2,24	0,4875	0,9875	0,0125	0,0325
2,25	0,4878	0,9878	0,0122	0,0317
2,26	0,4881	0,9881	0,0119	0,0310
2,27	0,4884	0,9884	0,0116	0,0303
2,28	0,4887	0,9887	0,0113	0,0297
2,29	0,4890	0,9890	0,0110	0,0290
2,30	0,4893	0,9893	0,0107	0,0283
2,31	0,4896	0,9896	0,0104	0,0277
2,32	0,4898	0,9898	0,0102	0,0270
2,33	0,4901	0,9901	0,0099	0,0264
2,34	0,4904	0,9904	0,0096	0,0258
2,35	0,4906	0,9906	0,0094	0,0252
2,36	0,4909	0,9909	0,0091	0,0246
2,37	0,4911	0,9911	0,0089	0,0241
2,38	0,4913	0,9913	0,0087	0,0235
2,39	0,4916	0,9916	0,0084	0,0229
2,40	0,4918	0,9918	0,0082	0,0224
2,41	0,4920	0,9920	0,0080	0,0219
2,42	0,4922	0,9922	0,0078	0,0213
2,43	0,4925	0,9925	0,0075	0,0208

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porțiunii mai mari (3)	Aria porțiunii mai mici (4)	*Y* pentru x/σ_x *(5)*
2,44	0,4927	0,9927	0,0073	0,0203
2,45	0,4929	0,9929	0,0071	0,0198
2,46	0,4931	0,9931	0,0069	0,0194
2,47	0,4932	0,9932	0,0068	0,0189
2,48	0,4934	0,9934	0,0066	0,0184
2,49	0,4936	0,9936	0,0064	0,0180
2,50	0,4938	0,9938	0,0062	0,0175
2,51	0,4940	0,9940	0,0060	0,0171
2,52	0,4941	0,9941	0,0059	0,0167
2,53	0,4943	0,9943	0,0057	0,0163
2,54	0,4945	0,9945	0,0055	0,0158
2,55	0,4946	0,9946	0,0054	0,0154
2,56	0,4948	0,9948	0,0052	0,0151
2,57	0,4949	0,9949	0,0051	0,0147
2,58	0,4951	0,9951	0,0049	0,0143
2,59	0,4952	0,9952	0,0048	0,0139
2,60	0,4953	0,9953	0,0047	0,0136
2,61	0,4955	0,9955	0,0045	0,0132
2,62	0,4956	0,9956	0,0044	0,0129
2,63	0,4957	0,9957	0,0043	0,0126
2,64	0,4959	0,9959	0,0041	0,0122
2,65	0,4960	0,9960	0,0040	0,0119
2,66	0,4961	0,9961	0,0039	0,0116
2,67	0,4962	0,9962	0,0038	0,0113
2,68	0,4963	0,9963	0,0037	0,0110
2,69	0,4964	0,9964	0,0036	0,0107
2,70	0,4965	0,9965	0,0035	0,0104
2,71	0,4966	0,9966	0,0034	0,0101
2,72	0,4967	0,9967	0,0033	0,0099
2,73	0,4968	0,9968	0,0032	0,0096
2,74	0,4969	0,9969	0,0031	0,0093

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porţiunii mai mari (3)	Aria porţiunii mai mici (4)	Y pentru x/σ_x (5)
2,75	0,4970	0,9970	0,0030	0,0091
2,76	0,4971	0,9971	0,0029	0,0088
2,77	0,4972	0,9972	0,0028	0,0086
2,78	0,4973	0,9973	0,0027	0,0084
2,79	0,4974	0,9974	0,0026	0,0081
2,80	0,4974	0,9974	0,0026	0,0079
2,81	0,4975	0,9975	0,0025	0,0077
2,82	0,4976	0,9976	0,0024	0,0075
2,83	0,4977	0,9977	0,0023	0,0073
2,84	0,4977	0,9977	0,0023	0,0071
2,85	0,4978	0,9978	0,0022	0,0069
2,86	0,4979	0,9979	0,0021	0,0067
2,87	0,4979	0,9979	0,0021	0,0065
2,88	0,4980	0,9980	0,0020	0,0063
2,89	0,4981	0,9981	0,0019	0,0061
2,90	0,4981	0,9981	0,0019	0,0060
2,91	0,4982	0,9982	0,0018	0,0058
2,92	0,4982	0,9982	0,0018	0,0056
2,93	0,4983	0,9983	0,0017	0,0055
2,94	0,4984	0,9984	0,0016	0,0053
2,95	0,4984	0,9984	0,0016	0,0051
2,96	0,4985	0,9985	0,0015	0,0050
2,97	0,4985	0,9985	0,0015	0,0048
2,98	0,4986	0,9986	0,0014	0,0047
2,99	0,4986	0,9986	0,0014	0,0046
3,00	0,4987	0,9987	0,0013	0,0044
3,01	0,4987	0,9987	0,0013	0,0043
3,02	0,4987	0,9987	0,0013	0,0042
3,03	0,4988	0,9988	0,0012	0,0040
3,04	0,4988	0,9988	0,0012	0,0039
3,05	0,4989	0,9989	0,0011	0,0038

Scor standard $z = x/\sigma_x$(1)	Aria de la medie la z (2)	Aria porţiunii mai mari (3)	Aria porţiunii mai mici (4)	Y pentru x/σ_x (5)
3,06	0,4989	0,9989	0,0011	0,0037
3,07	0,4989	0,9989	0,0011	0,0036
3,08	0,4990	0,9990	0,0010	0,0035
3,09	0,4990	0,9990	0,0010	0,0034
3,10	0,4990	0,9990	0,0010	0,0033
3,11	0,4991	0,9991	0,0009	0,0032
3,12	0,4991	0,9991	0,0009	0,0031
3,13	0,4991	0,9991	0,0009	0,0030
3,14	0,4992	0,9992	0,0008	0,0029
3,15	0,4992	0,9992	0,0008	0,0028
3,16	0,4992	0,9992	0,0008	0,0027
3,17	0,4992	0,9992	0,0008	0,0026
3,18	0,4993	0,9993	0,0007	0,0025
3,19	0,4993	0,9993	0,0007	0,0025
3,20	0,4993	0,9993	0,0007	0,0024
3,21	0,4993	0,9993	0,0007	0,0023
3,22	0,4994	0,9994	0,0006	0,0022
3,23	0,4994	0,9994	0,0006	0,0022
3,24	0,4994	0,9994	0,0006	0,0021
3,30	0,4995	0,9995	0,0005	0,0017
3,40	0,4997	0,9997	0,0003	0,0012
3,50	0,4998	0,9998	0,0002	0,0009
3,60	0,4998	0,9998	0,0002	0,0006
3,70	0,4999	0,9999	0,0001	0,0004

Notă: intrările din acest tabel au fost generate utilizând un program pe computer.

Bibliografie

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*, 12–15.

Acklin, M.W., McDowell, C. J., Verschell, M. S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment, 74*, 15–47.

Aiken, L. R. (1996). *Rating scales and checklists: Evaluating behavior, personality and attitudes.* New York: Wiley.

Aiken, L. R. (1997). *Questionnaires and inventories: Surveying opinions and assessing personality.* New York: Wiley.

American Association for Counseling and Development. (1988). *Responsibilities of users of standardized tests.*Washington, DC: Author.

American Counseling Association. (1995). *Code of ethics and standards of practice.* Alexandria, VA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (ed. a 4-a). Washington, DC: Author.

American Psychological Association. (1953). *Ethical standards of psychologists.*Washington, DC: Author.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.*Washington, DC: Author.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.*Washington, DC: Author.

American Psychological Association. (2000). *Report of the task force on test user qualifications.*Washington, DC: Author.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (ed. a 5-a). Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57*, 1060–1073.

American Psychological Association. (2003). *FAQ/Finding information about psychological tests.* Washington, DC: Author. Accesat la data de 5 iulie 2003 pe World Wide Web: http://www.apa.org/science/faq-findtests.html

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational & psychological tests.* Washington, DC: American Psychological Association.

American Psychological Association, Committee on Legal Issues. (1996). Strategies for private practitioners coping with subpoenas or compelled testimony for client records or test data. *Professional Psychology: Research and Practice, 27*, 245–251.

American Psychological Association, Committee on Psychological Tests and Assessment. (1996). Statement on the disclosure of test data. *American Psychologist, 51*, 644–648.

American Psychological Association, Committee on Psychological Tests and Assessment. (2003). *Statement on the use of secure psychological tests in the education of graduate and undergraduate psychology students.* Accesat la data de 19 februarie 2003, la adresa http://www.apa.org./science/securetests.html

Ames, L. B. (1989). *Arnold Gesell: Themes of his work.* New York: Human Sciences Press.

Anastasi, A. (1954). *Psychological testing.* New York: Macmillan.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

Anastasi, A. (1988). *Psychological testing* (ed. a 6-a). New York: Macmillan.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Ed.), *Test validity* (pp. 19–32). Hillsdale, NJ: Erlbaum.

Archer, R. P. (1987). *Using the MMPI with adolescents*. Hillsdale, NJ: Erlbaum.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). Accesat pe data de 10 mai 2003 la adresa http://ericae.net/irt [disponibil în prezent la http://edres.org/irt/baker/]

Balma, M. J. (1959). The concept of synthetic validity. *Personnel Psychology, 12*, 395–396.

Baugh, F. (2003). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 31–41). Thousand Oaks, CA: Sage.

Bell, D. E., Raiffa, H., & Tversky, A. (Ed.). (1988). *Decision-making: Descriptive, normative, and prescriptive interactions*. New York: Cambridge University Press.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice: Issues in constructed response, performance testing , and portfolio assessment*. Hillsdale, NJ: Erlbaum.

Beutler, L. E., & Groth-Marnat, G. (Ed.). (2003). *Integrative assessment of adult personality* (2nd ed.). New York: Guilford.

Blatt, J. (Producer/Writer/Director). (1989). *Against all odds: Inside statistics*. [videocasetă VHS]. (disponibilă la The Annenberg/CPB Project, 901 E St., NW, Washington, DC 2004–2006)

Bollen, K. A., & Long, J. S. (Ed.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Bondy, M. (1974). Psychiatric antecedents of psychological testing (before Binet). *Journal of the History of the Behavioral Sciences, 10*, 180–194.

Boring, E. G. (1950). *A history of experimental psychology* (ed. a 2-a). New York: Appleton-Century-Crofts.

Borum, R. (1996). Improving the clinical practice of violence risk assessment: Technology, guidelines, and training. *American Psychologist, 51*, 945–956.

Boudreau, J.W. (1991). Utility analysis for decisions in human resource management. În M. D. Dunnette & L. M. Hough (Ed.). *Handbook for industrial and organizational psychology* (ed. a 2-a,Vol. 2, pp. 621–745). Palo Alto, CA: Consulting Psychologists Press.

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist, 44*, 576–578.

Brannick, M. T., & Levine, E. L. (2002). *Job analysis: Methods, research, and applications for human resource management in the new millenium*. Thousand Oaks, CA: Sage.

Braun, H. L., Jackson, D. N., & Wiley, D. E. (2002). *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Erlbaum.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Briel, J. B., O'Neill, K., & Scheuneman, J. D. (Eds.). (1993). *GRE technical manual*. Princeton, NJ: Educational Testing Service.

Broadus, R. N., & Elmore, K. E. (1983). The comparative validities of undergraduate grade point average and of part scores on the Graduate Record Examinations in the prediction of two criterion measures in a graduate library school program. *Educational and Psychological Measurement, 43*, 543–546.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37*, 65–76.

Brown, C.W., & Ghiselli, E. E. (1953). Percent increase in proficiency resulting from use of selective devices. *Journal of Applied Psychology, 37*, 341–345.

Bryant, F. B., & Yarnold, P. R. (1995). Principal components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold, *Reading and understanding multivariate statistics* (pp. 99–136). Washington, DC: American Psychological Association.

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice, 33,* 148–154.

Buchwald, A. M. (1965). Values and the use of tests. *Journal of Consulting Psychology, 29,* 49–54.

Butcher, J. N. (2003). Computerized psychological assessment. În J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 141–163). Hoboken, NJ: Wiley.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory–2 (MMPI-2): Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Campbell, D. T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitraitmultimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organization psychology. În M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (ed. a 2-a, Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P. (1994). Alternative models of job performance and their implications for selection and classification. În M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 33–51). Hillsdale, NJ: Erlbaum.

Campbell, J. P., & Knapp, D. J. (Eds.). (2001). *Exploring the limits of personnel selection and classification.* Mahwah, NJ: Erlbaum.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: Cambridge University Press.

Carroll, J. B. (2002). The five-factor personality model: How complete and satisfactory is it? În H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 97–126). Mahwah, NJ: Erlbaum.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Erlbaum.

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (ed. a 3-a, pp. 201–219). New York: American Council on Education/Macmillan.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (ed. a 2-a). Hillsdale, NJ: Erlbaum.

Constantinople, A. (1973). Masculinity-femininity: An exception to a famous dictum? *Psychological Bulletin, 80,* 389–407.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual.* Odessa, FL: Psychological Assessment Resources.

Cowles, M. (2001). *Statistics in psychology: An historical perspective* (ed. a 2-a). Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1949). *Essentials of psychological testing.* New York: Harper & Row.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Cronbach, L. J. (1988). Five perspectives on validity argument. În H. Wainer & H. I. Braun (Ed.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Champaign: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

DeVellis, R. F. (2003). *Scale development: Theory and application* (ed. a 2-a). Thousand Oaks, CA: Sage.

Digman, J. M. (1990). Personality structure: Emergence of the Five-Factor Model. *Annual Review of Psychology, 41,* 417–440.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment*. Mahwah, NJ: Erlbaum.

DuBois, P. H. (1970). *A history of psychological testing*. Boston: Allyn & Bacon.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335–337.

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T.W., Meyer, G. J., & Moreland, K. (2000). Problems and limitations in the use of psychological assessment in the contemporary health care delivery system. *Professional Psychology: Research and Practice, 31,* 131–140.

Ekstrom, R. B., & Smith, D. K. (Eds.). (2002). *Assessing individuals with disabilities in educational, employment, and counseling settings*. Washington, DC: American Psychological Association.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8,* 341–349.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger, (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 1–15). Mahwah, NJ: Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Erikson, J., & Vater, S. (1995). *IDA administration manual: Procedures summary Provence Birth-to-Three Developmental Profile*. Itasca, IL: Riverside.

Eyde, L. E., Moreland, K. L., Robertson, G. J., Primoff, E. S., & Most, R. B. (1988). *Test user qualifications: A data-based approach to promoting good test use*.Washington, DC: American Psychological Association.

Fancher, R. E. (1985). *The intelligence men: Makers of the IQ controversy*. New York: Norton.

Fancher, R. E. (1996). *Pioneers of psychology* (ed. a 3-a). New York: Norton.

Fink, A. (2002). *How to ask survey questions* (ed. a 2-a, Vol. 2). Thousand Oaks, CA: Sage.

Finn, S. E., & Kamphuis, J. H. (1995). What a clinician needs to know about base rates. În J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp. 224–235). New York: Oxford University Press.

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9,* 374–385.

Fischer, C. T. (1994). *Individualizing psychological assessment*. Hillsdale, NJ: Erlbaum. (publicată inițial în 1985)

Fischer, C. T. (2000). Collaborative, individualized assessment. *Journal of Personality Assessment, 74,* 2–14.

Flanagan, S. G. (1992). Review of the Whitaker Index of Schizophrenic Thinking. *The Eleventh Mental Measurements Yearbook,* pp. 1033–1034.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains. *Psychological Bulletin, 95,* 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Franzen, M. D. (2000). *Reliability and validity in neuropsychological assessment* (2nd ed.). New York: Kluwer Academic/Plenum.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.). (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Erlbaum.

Frisch, M. B. (1994). *QOLI: Quality of Life Inventory.* Minneapolis, MN: National Computer Systems.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519–521.

Goldberg, E. L., & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement, 52,* 1019–1027.

Goldman, L. (1971). *Using tests in counseling* (2nd ed.). New York: Appleton-Century-Crofts.

Goldman, B. A., Mitchell, D. F., & Egelson, P. (Eds.). (1997). *Directory of unpublished experimental mental measures: Volume 7 (1990–1995).* Washington, DC: American Psychological Association.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Assessment of aphasia and related disorders* (ed. a 3-a). Baltimore, MD: Lippincott, Williams & Wilkins.

Gronlund, N. E. (2003). *Assessment of student achievement* (ed. a 7-a). Boston: Allyn & Bacon.

Groth-Marnat, G. (1997). *Handbook of psychological assessment* (ed. a 3-a). New York: Wiley.

Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (ed. a 2-a, Vol. 2, pp. 327– 397). Palo Alto, CA: Consulting Psychologists Press.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Erlbaum.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gustafsson, J.-E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73–95). Mahwah, NJ: Erlbaum.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking.* Boston: Allyn & Bacon.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (ed. a 2-a). Mahwah, NJ: Erlbaum.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52,* 1091–1101.

Hambleton, R. K., & Rogers, H. J. (1991). Advances in criterion-referenced measurement. În R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 3–43). Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Handler, L., & Meyer, G. J. (1998). The importance of teaching and learning personality assessment. În L. Handler & M. J. Hilsenroth (Eds.), *Teaching and learning personality assessment* (pp. 3–30). Mahwah, NJ: Erlbaum.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hathaway, S. R., & McKinley, J. C. (1940). A Multiphasic Personality Schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10,* 249–254.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41–45.

Hersen, M., & Van Hasselt, V. B. (Eds.). (1998). *Basic interviewing: A practical guide for counselors and clinicians.* Mahwah, NJ: Erlbaum.

Holland, P.W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. În H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Howell, D. C. (2002). *Statistical methods for psychology* (ed. a 5-a). Pacific Grove, CA: Duxbury.

Hummel, T. J. (1999). The usefulness of tests in clinical decisions. În J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 59–112). Boston: Allyn & Bacon.

Hunter, J. E., & Schmidt, F. L. (1981). Fitting people into jobs: The impact of personnel selection on national productivity. În M. A. Dunnette & E. A. Fleishman (Eds.), *Human performance and productivity: Vol. 1. Human capability assessment* (pp. 233–284). Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulations: The critical role of meta-analysis. *Psychology, Public Policy, and Law, 2*, 324–347.

Impara, J. C., & Plake, B. S. (Eds.). (1998). *The Thirteenth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute.

International Test Commission. (2000). *International guidelines for test use: Version 2000*. Accesată la 8 iunie 2002, la adresa http://www.intestcom.org

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kormen, E., & Kurland, L. T. (1992). Mayo's older Americans normative studies: WAIS-R norms for ages 56 to 97. *Clinical Neuropsychologist, 6*(Suppl.), 1–30.

Jaeger, R. M. (1989). Certification of student competence. În R. L. Linn (Ed.), *Educational measurement* (ed. a 3-a, pp. 485–514). New York: American Council on Education/Macmillan.

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin, 80*, 75–83.

James, W. (1890). *The principles of psychology* (Vols. 1–2). New York: Henry Holt.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285–306.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Johnson, S. (2000). *Taking the anxiety out of taking tests: A step-by-step guide*. New York: Barnes & Noble Books.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*.Washington, DC: Author.

Joint Committee on Testing Practices. (1998). *Rights and responsibilities of test takers: Guidelines and expectations*. Accesată la 12 iulie 2003 la http://www.apa.org/science/ttrr.html

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.

Kamphaus, R.W. (2001). *Clinical assessment of child and adolescent intelligence* (ed. a 2-a). Boston: Allyn & Bacon.

Kamphaus, R.W., & Frick, P. J. (2002). *Clinical assessment of child and adolescent personality and behavior* (ed. a 2-a). Boston: Allyn & Bacon.

Kaufman, A. S. (2000). Tests of intelligence. În R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 445–476). New York: Cambridge University Press.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test: Manual*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Lichtenberger, E. O. (2002). *Assessing adolescent and adult intelligence* (ed. a 2-a). Boston: Allyn & Bacon.

Kellogg, C. E., & Morton, N.W. (1999). *Beta III: Manual*. San Antonio, TX: The Psychological Corporation.

Kirk, R. E. (1999). *Statistics: An introduction* (4th ed.). Fort Worth, TX: Harcourt Brace.

Kirsch, I., & Sapirstein, G. (1998, June 26). Listening to Prozac but hearing placebo: A metaanalysis of antidepressant medication. *Prevention & Treatment, 1*, Article 0002a. Accesată la 2 iulie 1998 la http://journals.apa.org/prevention/volume 1/pre0010002a.html

Kobrin, J. L., Camara, W. J., & Milewski, G. B. (2002). *The utility of the SAT I and SAT II for admissions decisions in California and the nation.* (College Board Research Report No. 2002–6). New York: College Entrance Examination Board.

Kubiszyn, T.W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., & Eisman, E. J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice, 31*, 119–130.

Kuncel, N. R., Campbell, J. P., & Ones, D. S. (1998). Validity of the Graduate Record Examination: Estimated or tacitly known? *American Psychologist, 53*, 567–568.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162–181.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 11*, 1183–1192.

Lanyon, R. I., & Goodstein, L. D. (1997). *Personality assessment* (ed. a 3-a). New York: Wiley.

Lees-Haley, P. R. (1996). Alice in Validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51*, 981–983.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York: Farrar, Straus & Giroux.

Lenney, E. (1991). Sex roles: The measurement of masculinity, femininity, and androgyny. În J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 573–660). San Diego, CA: Academic Press.

Lezak, M. D. (1995). *Neuropsychological assessment* (ed. a 3-a). New York: Oxford University Press.

Lichtenberg, J.W., & Goodyear, R. K. (Eds.). (1999). *Scientist-practitioner perspectives on test interpretation*. Boston: Allyn & Bacon.

Linn, R. L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice, 13*, 12–14.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues & Practice, 17*, 28–30.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Upper Saddle River, NJ: Merrill.

Livesley, W. J. (Ed.). (2001). *Handbook of personality disorders: Theory, research, and treatment*. New York: Guilford.

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph Supplement]. *Psychological Reports, 3*, 635–694.

Lowman, R. L. (1991). *The clinical practice of career assessment: Interests, abilities, and personality.*Washington, DC: American Psychological Association.

Lubinski, D., & Dawis, R. V. (1992). Aptitudes, skills, and proficiencies. În M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organization psychology* (ed. a 2-a, Vol. 3, pp. 1–59).

Maloney, M. P., & Ward, M. P. (1976). *Psychological assessment: A conceptual approach*. New York: Oxford University Press.

Marston, A. R. (1971). It is time to reconsider the Graduate Record Examination. *American Psychologist, 26*, 653–655.

Maruish, M. E. (Ed.). (2004). *The use of psychological testing for treatment planning and outcome assessment* (3rd ed., Vols. 1–3). Mahwah, NJ: Erlbaum.

Masling, J. (1960). The influence of situational and interpersonal variables in projective testing. *Psychological Bulletin, 57,* 65–85.

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52,* 443–451.

McClelland, D. C. (1958). Methods of measuring human motivation. În J. W. Atkinson (Ed.). *Motives in fantasy, action, and society* (pp. 7–45). Princeton, NJ: Van Nostrand.

McReynolds, P. (1986). History of assessment in clinical and educational settings. În R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 42–80). New York: Guilford.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil tests: A meta-analysis. *Psychological Bulletin, 114,* 449–458.

Meehl, P. E. (1956). Wanted: A good cookbook. *American Psychologist, 11,* 263–272.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11,* 3–9, 20.

Merriam-Webster's collegiate dictionary (ed. a 10-a). (1995). Springfield, MA: Merriam-Webster.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. În H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Meyer, G. J. (1992). Response frequency problems in the Rorschach: Clinical and research implications with suggestions for the future. *Journal of Personality Assessment, 58,* 231–244.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Millon, T., Millon, C., & Davis, R. (1994). *MCMI-III manual: Millon Clinical Multiaxial Inventory–III.* Minneapolis, MN: National Computer Systems.

Mitchell, R. (1992). *Testing for learning.* New York: Free Press/Macmillan.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R., & Heilbronner, R. L. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale–Revised. *Professional Psychology: Research & Practice, 26,* 491–498.

Monahan, J., & Steadman, H. J. (2001). Violence risk assessment: A quarter century of research. În L. E. Frost & R. J. Bonnie (Eds.), *The evolution of mental health law* (pp. 195–211). Washington, DC: American Psychological Association.

Moreland, K. L. (1991). Assessment of validity in computer-based test interpretations. În T. B. Gutkin & S. L. Wise (Eds.), *The computer and the decision-making process* (pp. 43–74). Hillsdale, NJ: Erlbaum.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55,* 309–316.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23,* 5–12.

Murphy, K. R., Fleishman, E. A., & Cleveland, J. N. (Eds.). (2003). *Validity generalization: A critical review.* Mahwah, NJ: Erlbaum.

Murphy, L. L., Plake, B. S., Impara, J. C., & Spies, R. A. (Eds.). (2002). *Tests in print VI: An index to tests, test reviews, and the literature on specific tests*. Lincoln, NE: Buros Institute.

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59*, 150–162.

National Academy of Neuropsychology. (1999, May 15). *Policy statement: Presence of third party observers during neuropsychological testing*. Accesată la data de 3 august 2003 la adresa http://nanonline.org/content/text/paio/thirdparty.shtm

National Association of School Psychologists. (2000). *Professional conduct manual: Principles for professional ethics guidelines for the provision of school psychological services*. Accesată la data de 27 august 2003, la adresa http://www.nasponline.org/pdf/ProfessionalCond.pdf

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*.Washington, DC: American Psychological Association.

Nevo, B., & Jäger, R. S. (Eds.). (1993). *Educational and psychological testing: The test taker's outlook*. Toronto: Hogrefe & Huber.

Newman, E. (1996). *No more test anxiety: Effective steps for taking tests and achieving better grades*. Los Angeles: Learning Skills Publications.

O'Leary, B. J., & Norcross, J. C. (1998). Lifetime prevalence of mental disorders in the general population. În G. P. Koocher, J. C. Norcross, & S. S. Hill III (Eds.), *Psychologists' desk reference* (pp. 3–6). New York: Oxford University Press.

Onwuegbuzie, A. J., & Daniel, L. G. (2002). A framework for reporting and interpreting internal consistency reliability estimates. *Measurement and Evaluation in Counseling and Development, 35*, 89–103.

Ownby, R. L. (1997). *Psychological reports: A guide to report writing in professional psychology* (ed. a 3-a). New York: Wiley.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. În R. L. Linn (Ed.), *Educational measurement* (ed. a 3-a, pp. 221–262). New York: American Council on Education/Macmillan.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1–9.

Primi, R. (2002, August). *Contribution to the development of a fluid intelligence scale*. Poster prezentat la întâlnirea anuală a American Psychological Association, Chicago, IL.

Primoff, E. S. (1959). Empirical validations of the J-coefficient. *Personnel Psychology, 12*, 413–418.

Primoff, E. S., & Eyde, L. D. (1988). Job element analysis. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol. 2, pp. 807–824). New York: Wiley.

Provence, S., Erikson, J., Vater, S., & Palmeri, S. (1995). *Infant-Toddler Developmental Assessment (IDA)*. Itasca, IL: Riverside.

Psychological Corporation. (1997). *WAIS-III/WMS-III technical manual*. San Antonio, TX: Author.

Ramsay, M. C., & Reynolds, C. R. (2000). Development of a scientific test: A practical guide. În G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (ed. a 3-a, pp. 21–42). Oxford, UK: Elsevier.

Randhawa, B. S. (1992). Review of the Clerical Abilities Battery. *The Eleventh Mental Measurements Yearbook*, pp. 187–189.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press. (Lucrare publicată inițial în 1960)

Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Erlbaum.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93–103.

Robertson, G. J. (1992). Psychological tests: Development, publication, and distribution. În M. Zeidner & R. Most (Eds.), *Psychological testing: An inside view* (pp. 169–214). Palo Alto, CA: Consulting Psychologists Press.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1991). *Measures of personality and social psychological attitudes*. San Diego, CA: Academic Press.

Rogers, R. (1997). *Clinical assessment of malingering and deception* (ed. a 2-a). New York: Guilford Press.

Rogers, T. B. (1995). *The psychological testing enterprise: An introduction*. Pacific Grove, CA: Brooks/Cole.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80* (1, Whole No. 609).

Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research, & Evaluation, 4*(2). Accesată la data de 9 august 2003, la adresa http://edresearch.org/pare/getvn.asp?v=4&n=2

Russell, D.W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in *Personality and Social Psychology Bulletin. Personality and Social Psychology Bulletin, 28*, 1629–1646.

Sacks, P. (1999). *Standardized minds: The high price of America's testing culture and what we can do to change it*. Cambridge, MA: Perseus Books.

Sapp, M. (1999). *Test anxiety: Applied research, assessment, and treatment* (2nd ed.). Latham, MD: University Press of America.

Sandoval, J., Frisby, C. L., Geisinger, K. F., Scheuneman, J. D., & Grenier, J. R. (1998). *Test interpretation and diversity: Achieving equity in assessment*.Washington, DC: American Psychological Association.

Sarason, I. G. (Ed.). (1980). *Test anxiety: Theory, research, and applications*. Hillsdale, NJ: Erlbaum.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (ed. a 4-a). San Diego, CA: Author.

Sattler, J. M. (2002). *Assessment of children: Behavioral and clinical applications* (ed. a 4-a). San Diego, CA: Author.

Sawilowsky, S. S. (2003). Reliability: Rejoinder to Thompson and Vacha-Haase. În B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 149–154). Thousand Oaks, CA: Sage.

Schaie, K.W. (1994). The course of adult intellectual development. *American Psychologist, 49*, 304– 313.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T.W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*, 609–626.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalization methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78*, 3–12.

Scruggs, T. E., & Mastropieri, M. A. (1992). *Teaching test-taking skills: Helping students show what they know*. Cambridge, MA: Brookline Books.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Silverlake, A. C. (1999). *Comprehending test manuals: A guide and workbook*. Los Angeles: Pyrczak.

Smith, C. P. (1992). Reliability issues. In C. P. Smith (Ed.) et al., *Motivation and personality: Handbook of thematic content analysis* (pp. 126–139). New York: Cambridge University Press.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (ed. a 4-a). Accesată la data de 14 noiembrie 2003 la adresa http://www.siop.org/_Principles/principles.pdf

Spearman, C. (1904a). „General intelligence", objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Spearman, C. (1904b). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72–101.

Spence, J. T. (1993). Gender-related traits and gender ideology evidence for a multifactorial theory. *Journal of Personality and Social Psychology, 64,* 624–635.

Spence, J. T., & Helmreich, R. L. (1972). The Attitudes Toward Women Scale: An objective instrument to measure attitudes toward the rights and roles of women in contemporary society. *JSAS Catalog of Selected Documents in Psychology, 2,* 66–67.

Spielberger, C. D., & Vagg, P. R. (Eds.). (1995). *Test anxiety: Theory, assessment, and treatment.*Washington, DC: Taylor & Francis.

Steele, C. M. (1997). A threat in thin air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

Sternberg, R. J. (1986). *Intelligence applied: Understanding and increasing your intellectual skills.* San Diego, CA: Harcourt Brace Jovanovich.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? A case study. *American Psychologist, 52,* 630–641.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (Eds.). (1994). *Biodata handbook: Theory, research, and use of biographical information for selection and performance prediction.* Palo Alto, CA: Consulting Psychologists Press.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.

Tallent, N. (1993). *Psychological report writing* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. Discussion and tables. *Journal of Applied Psychology, 23,* 565– 578.

Tenopyr, M. L. (1986). Needed directions for measurement in work settings. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 269–288). Hillsdale, NJ: Erlbaum.

Terrill, D. R., Friedman, D. G., Gottschalk, L. A., & Haaga, D. A. (2002). Construct validity of the Life Orientation Test. *Journal of Personality Assessment, 79,* 550–563.

Test security: Protecting the integrity of tests. (1999). *American Psychologist, 54,* 1078.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring.* Mahwah, NJ: Erlbaum.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 25–32.

Thompson, B. (2003a). A brief introduction to generalizability theory. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 43–58). Thousand Oaks, CA: Sage.

Thompson, B. (Ed.). (2003b). *Score reliability: Contemporary thinking on reliability issues.* Thousand Oaks, CA: Sage.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16,* 433–451.

Turner, S. M., DeMers, S. T., Fox, H. R., & Reed, G. M. (2001). APA's guidelines for test user qualifications: An executive summary. *American Psychologist, 56,* 1099–1113.

U.S. Department of Education, Office for Civil Rights. (2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy-makers.* Washington, DC: Author. Accesată la data de 5 aprilie 2001 la adresa http://www.ed.gov/offices/OCR/testing/index1.html

Urbina, S., & Ringby, K. C. (2001, August). *Sex differences in mechanical aptitude.* Poster prezentat la întâlnirea anuală a American Psychological Association, San Francisco, CA.

Urdan, T. C. (2001). *Statistics in plain English.* Mahwah, NJ: Erlbaum.

Uzgiris, I. C., & Hunt, J. McV. (1975). *Assessment in infancy: Ordinal Scales of Psychological Development.* Urbana, IL: University of Illinois Press.

Viglione, D. J., & Rivera, B. (2003). Assessing personality and psychopathology with projective methods. In I. B. Weiner (Series Ed.). & J. R. Graham & J. A. Naglieri (Vol. Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 531–552). Hoboken, NJ: Wiley.

Vogt, W. P. (1998). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (ed. a 2-a). Thousand Oaks, CA: Sage.

von Mayrhauser, R. T. (1992). The mental testing community and validity: A prehistory. *American Psychologist, 47,* 244–253.

Wainer, H. (2000). CATs: Whither and whence. *Psicologica, 21,* 121–133.

Wainer, H. (cu Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L. & Thissen, D.). (2000). *Computer adaptive testing: A primer* (ed. a 2-a). Mahwah, NJ: Erlbaum.

Wallace, S. R. (1965). Criteria for what? *American Psychologist, 20,* 411–417.

Wechsler, D. (1939). *Wechsler-Bellevue Intelligence Scale.* New York: The Psychological Corporation.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children (WISC): Manual.* New York: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children–Third Edition (WISC-III): Manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–Third Edition (WAIS-III): Administration and scoring manual.* San Antonio, TX: The Psychological Corporation.

Whitaker, L. (1980). *Objective measurement of schizophrenic thinking: A practical and theoretical guide to the Whitaker Index of Schizophrenic Thinking.* Los Angeles: Western Psychological Services.

Wigdor, A. K., & Green, B. F., Jr. (1991). (Eds.), *Performance assessment for the workplace: Vol. 2. Technical issues.* Washington, DC: National Academy Press.

Wiggins, J. S., & Pincus, A. L. (1992). Personality: Structure and assessment. *Annual Review of Psychology, 43,* 493–504.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Willingham, W.W., Pollack, J. M., & Lewis, C. (2000). *Grades and test scores: Accounting for observed differences* (ETS RR-00-15). Princeton, NJ: Educational Testing Service.

Young, J.W. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review.* (College Board Research Report No. 2001–6). New York: College Entrance Examination Board.

Zeidner, M. (1998). *Test anxiety: The state of the art.* New York: Plenum.

Ziskin, J. (1995). *Coping with psychiatric and psychological testimony* (ed. a 5-a, Vol. 1–3). Los Angeles: Law and Psychology Press.

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education.* New York: Routledge Falmer.

Mulțumiri

Autoarea este recunoscătoare lui Dr. Joan Hill și Dr. Ping Sa, pentru corectura și comentariile la manuscris și Dr. Bill Wilson pentru ajutorul dat în construirea tabelului din Anexa C. Mulțumiri speciale pentru David Wilson de la Center for Instruction and Research Technology, University of North Florida pentru pregătirea celor mai multe dintre figurile prezente în această lucrare și lui Tracey Belmont de la John Wiley & Sons pentru sprijinul acordat în fiecare fază a acestei cărți.

PSIHOLOGIE

Testarea psihologică subliniază principiile de bază ale psihometriei şi oferă informaţiile necesare pentru înţelegerea, evaluarea şi utilizarea competentă a testelor psihologice. După o introducere în istoria dezvoltării şi utilizării testelor psihologice, autoarea prezintă metodele statistice folosite cel mai frecvent în testare, cadrul de referinţă pentru interpretarea scorurilor unui test, consideraţii legate de fidelitate, validitate şi construirea itemilor unui test, precum şi indicaţii cu privire la selectarea, administrarea şi scorarea testelor şi comunicarea rezultatelor obţinute. În fiecare capitol sunt subliniate conceptele cheie, sunt indicate resurse suplimentare relevante. Teoria şi cercetarea sunt îmbinate echilibrat în text, iar materialul ilustrativ este din plin prezent. Cartea este utilă atât studenţilor la psihologie, cât şi practicienilor din domeniul evaluării psihologice, şi este o lectură accesibilă pentru un cititor nespecialist, interesat de domeniul testării psihologice.

SUSANA URBINA este profesor de psihologie la University of North Florida, unde predă din 1976 cursuri de testare psihologică şi evaluare psihologică. Este coautor al celei de-a şaptea ediţii a Psychological Testing, alături de Anne Anastasi. A publicat numeroase articole în domeniul testării psihologice, a condus cercetări de validare a testelor şi a fost consultant pentru proiecte de construire şi standardizare a diferitelor teste. Este preşedinta Comisiei pentru teste psihologice şi evaluare şi a Comisiei pentru practică şi standarde profesionale din cadrul Asociaţiei Americane de Psihologie (APA).

de Psihologie (APA).